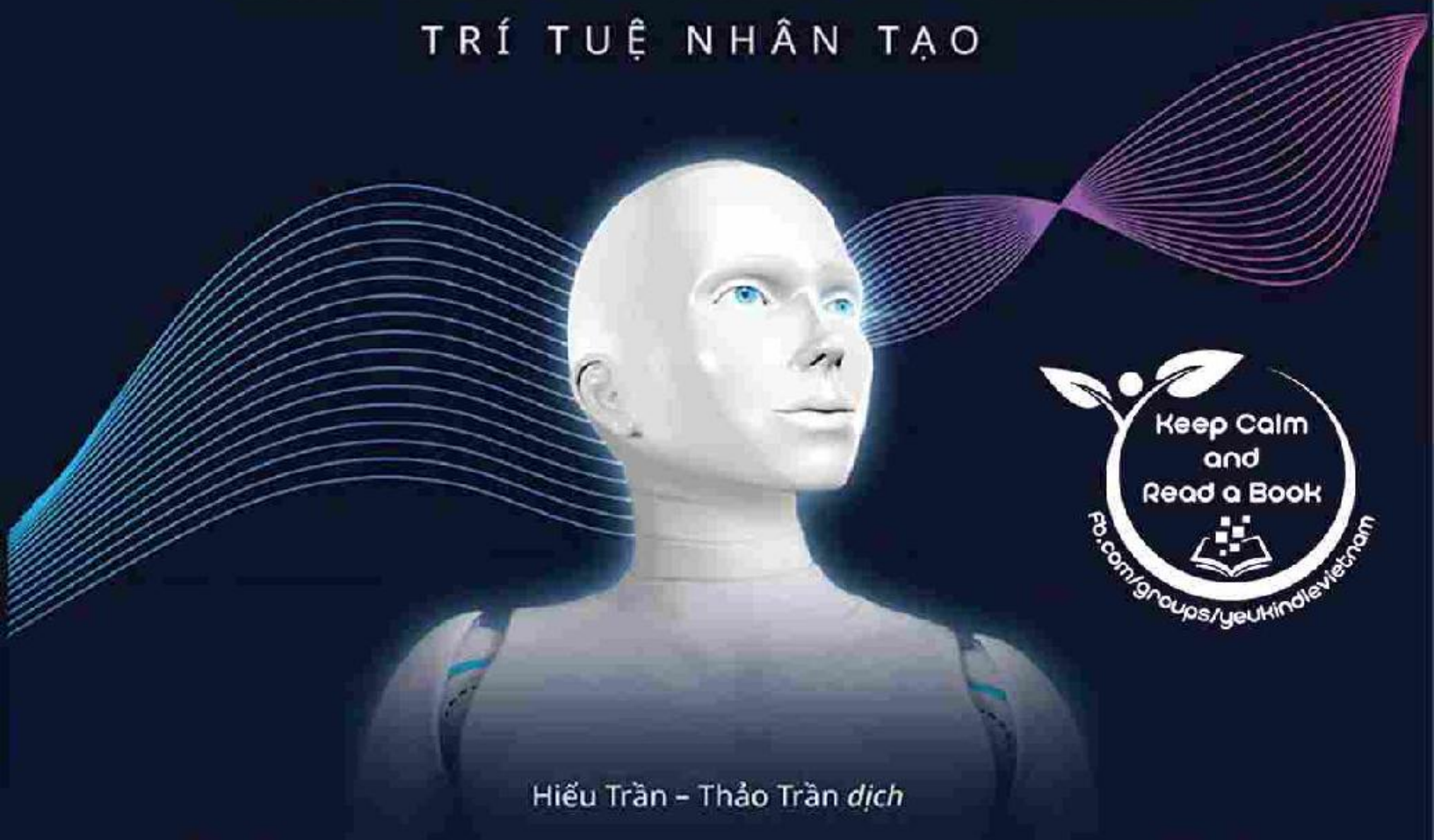


LIFE

3.0

LOÀI NGƯỜI TRONG KỶ NGUYÊN
TRÍ TUỆ NHÂN TẠO



Hiếu Trần - Thảo Trần dịch

MAX TEGMARK



EDUCATION
TECHNOLOGY
SCIENCE

alphabooks®

NHÀ XUẤT BẢN
THẾ GIỚI

THÔNG TIN EBOOK

Kindle Kobo Vietnam

www.facebook.com/groups/yeukindlevietnam

Keep Calm
and
Read a Book



Tên sách

Life 3.0 - Con người trong kỷ nguyên trí tuệ nhân tạo

Tác giả

Max Tegmark

Thể loại

Khoa học - Công nghệ

Nhà Xuất Bản

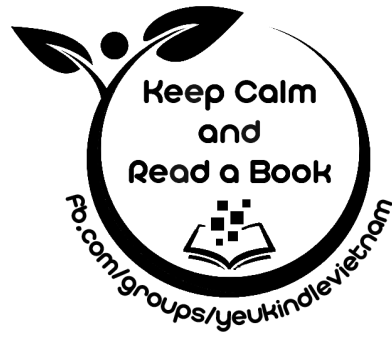
NXB Thế Giới

Năm

2020 kindlekobovn

—★—

Lưu ý: Cuốn này không có sách giấy để so sánh nên không làm được footnote chuẩn



Lời tựa

Tôi thực lòng biết ơn những người đã khích lệ và giúp đỡ tôi viết cuốn sách này, bao gồm:

gia đình, bạn bè, thầy cô, đồng nghiệp và cộng sự vì đã ủng hộ và tạo cảm hứng cho tôi suốt những năm qua;

Mẹ tôi vì đã khơi thắp sự tò mò trong tôi về ý thức và ý nghĩa;

Cha tôi vì tinh thần chiến đấu cho một thế giới tốt đẹp hơn;

các con trai tôi, Philip và Alexander, vì đã cho tôi thấy sự kỳ diệu của các trí-tuệ-mức-con-người đang phát triển;

tất cả những người tâm huyết với khoa học và công nghệ trên khắp thế giới đã đóng góp những câu hỏi, bình luận và động viên khuyến khích tôi trong nhiều năm qua để theo đuổi và xuất bản các ý tưởng của mình;

đại diện của tôi, John Brockman, vì đã dai dẳng thuyết phục cho tới khi tôi đồng ý viết sách này;

Bob Penna, Jesse Thaler và Jeremy England vì những cuộc thảo luận bổ ích về chuẩn tinh, cực đại nội tại sphaleron và nhiệt động học;

những người đã phản hồi về nhiều phần bản thảo của tôi, gồm có Mẹ, em trai tôi Per, Luisa Bahet, Rob Bensinger, Katerina Bergström, Erik Brynjolfsson, Daniela Chita, David Chalmers, Nima Deghani, Henry Lin, Elin Malmsköld, Toby Ord, Jeremy Owen, Lucas Perry, Anthony Romero, Nate Soares và Jaan Tallinn;

những siêu anh hùng đã bình luận cho các bản nháp cuốn sách, đó là Meia, Cha, Anthony Aguirre, Paul Almond, Matthew Graves, Phillip Helbig, Richard Mallah, David Marble, Howard Messing, Luiño Seoane, Marin Soljačić, biên tập viên Dan Frank và, hơn hết là,

Meia, nàng thơ yêu dấu và người bạn đồng hành của tôi; nếu không có sự ủng hộ, động viên và truyền cảm hứng bất tận của cô ấy, cuốn sách này sẽ không ra đời.

Phần mở đầu

Câu chuyện về nhóm Omega

Nhóm Omega là linh hồn của công ty. Trong khi cả tập đoàn bận kiếm tiền duy trì hoạt động với các ứng dụng trí tuệ nhân tạo (AI¹) hẹp, thì chỉ nhóm Omega dốc thân vào công cuộc chinh phục giấc mơ bấy lâu của ngài CEO: xây dựng trí tuệ nhân tạo tổng quát. Hầu hết các nhân viên khác đều coi “hội Omega” – tên gọi thân thiết của nhóm – như một cơn mơ mộng viễn vông, luôn luôn cách mục tiêu họ đề ra vài thập kỷ. Nhưng họ vẫn vui vẻ chiều chuộng nhóm, bởi họ thích uy thế mà công ty nhận được từ những công trình tối tân của hội Omega và họ cũng đánh giá cao những thuật toán cải tiến mà hội thi thoảng vẫn cung cấp.

Điều mọi người không nhận ra ở đây là hội Omega đã cẩn trọng tạo dựng hình tượng để che giấu một bí mật: họ chẳng còn cách bước chinh phục kế hoạch táo bạo nhất lịch sử loài người bao xa. Ngài CEO lôi cuốn của tập đoàn đã lựa chọn kỹ càng từng người một, không chỉ vì họ là những nhà nghiên cứu sáng giá, mà còn bởi họ có tham vọng, lý tưởng cũng như rất tận tụy trong việc giúp đỡ nhân loại. Ông đã nhắc nhở họ rằng kế hoạch này cực kỳ nguy hiểm và nếu các chính phủ đầy quyền lực biết được, họ sẽ làm gần như mọi thứ để ngăn chặn nhóm – bao gồm cả bắt cóc – hoặc, nếu được thì ăn cắp mã nguồn. Nhưng tất cả vẫn tham gia 100%, cũng chính vì lý do mà các nhà vật lý học hàng đầu thế giới tham gia vào Dự án Manhattan để phát triển vũ khí hạt nhân: họ cho rằng nếu họ không làm điều đó trước, một kẻ thiếu lý tưởng hơn sẽ làm.

Prometheus, AI mà họ xây dựng, đang ngày một giỏi hơn. Mặc dù khả năng nhận thức của nó còn kém xa con người trong nhiều lĩnh vực, chẳng hạn các kỹ năng xã hội, nhưng hội Omega đã nỗ lực hết sức để nó vượt trội ở một nhiệm vụ chính: lập trình các hệ thống AI khác. Họ đã lựa chọn chiến thuật này bởi họ tin vào thuyết bùng nổ trí tuệ của nhà toán học người Anh Irving Good vào năm 1965: “Hãy định nghĩa cỗ máy siêu thông minh là một cỗ máy có thể vượt xa mọi

hoạt động trí tuệ của con người, bất kể họ thông minh cỡ nào. Bởi thiết kế máy móc chính là một hoạt động trí tuệ, một cỗ máy siêu thông minh sẽ thiết kế được những cỗ máy thậm chí còn tốt hơn thế; như vậy chắc chắn sẽ xảy ra hiện tượng bùng nổ trí tuệ, và trí tuệ loài người sẽ bị bỏ xa. Do đó, cỗ máy siêu thông minh đầu tiên sẽ là thứ cuối cùng mà loài người cần phát minh, với điều kiện cỗ máy đó đủ ngoan ngoãn để cho chúng ta biết cách kiểm soát nó.”

Họ luận ra rằng, nếu có thể vận hành chu trình tự cải tiến đệ quy² này, cỗ máy sẽ sớm trở nên đủ thông minh để có thể tự dạy chính nó tất cả các kỹ năng hữu ích khác của con người.

Hàng triệu đô-la đầu tiên

Chín giờ sáng thứ Sáu, họ quyết định khởi động. Prometheus đang chạy rì rì trong cụm máy tính được tạo riêng cho nó, chúng nằm trên hàng dài các giá đỡ, trong một căn phòng điều hòa rộng, được kiểm soát ra vào. Vì lý do an toàn, Prometheus bị ngắt kết nối hoàn toàn với mạng Internet, nhưng nó có bản sao của phần lớn các trang web (Wikipedia, Thư viện Quốc hội, Twitter, một bộ sưu tập chọn lọc từ YouTube, phần lớn Facebook, v.v) để dùng làm dữ liệu học tập.³ Họ đã chọn thời điểm này để không bị quấy rầy: gia đình và bạn bè nghĩ họ đang trong kỳ nghỉ cuối tuần cùng tập đoàn. Khu bếp chứa đầy đồ ăn có thể bỏ lò vi sóng cùng nước uống tăng lực; họ đã sẵn sàng khởi hành.

Lúc mới khởi động, Prometheus vẫn kém họ một chút ở khoản lập trình hệ thống AI, nhưng bù lại, nó cực kỳ nhanh chóng trong việc giải quyết các vấn đề có thể khiến một người mất hàng nghìn năm, ngay trong lúc nhóm còn đang giải quyết mấy lon Bò Húc. Tới mười giờ sáng, nó đã tự thiết kế lại chính nó lần đầu tiên thành phiên bản 2.0, tốt hơn một chút nhưng vẫn kém con người. Thế nhưng, khi Prometheus 5.0 khởi động vào hai giờ chiều, hội Omega đã rất ngỡ ngàng: nó thổi bay mọi tiêu chuẩn đánh giá của họ và tốc độ tiến bộ có vẻ ngày càng nhanh hơn. Tới chập tối, họ quyết định tạm dừng Prometheus 10.0 để thực hiện giai đoạn 2 của kế hoạch: kiếm tiền.

Đối tượng đầu tiên của họ là MTurk – Amazon Mechanical Turk⁴. Sau khi triển khai vào năm 2005, khu chợ dành cho đóng góp cộng đồng qua Internet này đã phát triển nhanh chóng, với hàng vạn người ẩn danh trên thế giới cạnh tranh suốt ngày đêm để thực hiện các công việc lặt vặt có yêu cầu chi tiết gọi là HITs, “công việc đòi hỏi trí tuệ con người”⁵. Các công việc này bao gồm từ việc chép lại các bản ghi âm tới phân loại ảnh và viết mô tả cho trang web, và tất cả đều có một điểm chung: nếu bạn làm tốt thì sẽ không ai biết bạn là một AI. Prometheus 10.0 có thể hoàn thành khá tốt khoảng một nửa các nhóm việc. Với mỗi nhóm như vậy, hội Omega đều để Prometheus thiết kế riêng một mô-đun phần mềm AI hợp tình gọn nhằm thực hiện chỉ riêng nhóm việc đó. Sau đó họ sẽ tải mô-đun này lên Amazon Web Services, một nền tảng điện toán đám mây vận hành trên mọi máy ảo mà họ có thể thuê. Với mỗi đô-la trả cho bộ phận điện toán đám mây của Amazon, họ kiếm được hơn hai đô-la từ MTurk của Amazon. Và Amazon chẳng hề ngờ rằng cơ hội kiếm lời chênh lệch tuyệt vời như vậy lại tồn tại trong tập đoàn của chính họ!

Để che giấu hoạt động, họ đã bí mật tạo ra hàng nghìn tài khoản MTurk từ nhiều tháng trước bằng tên giả, và giờ các mô-đun do Prometheus xây dựng sẽ sử dụng danh tính đó. Khách hàng MTurk thường thanh toán sau khoảng tám tiếng, khi đó hội Omega sẽ tái đầu tư số tiền kiếm được để kéo dài thời gian dùng điện toán đám mây cho các mô-đun nhiệm vụ ngày càng tốt hơn do phiên bản tự tiến bộ mới nhất của Prometheus tạo thành. Bởi cứ sau tám tiếng lại có thể nhân đôi số tiền, họ nhanh chóng cung cấp khả năng vượt quá yêu cầu từ MTurk và nhận ra rằng không thể kiếm nhiều hơn một triệu đô-la một ngày mà không gây ra những chú ý không cần thiết. Nhưng như vậy cũng thừa đủ vốn cho bước tiếp theo của họ và tránh được việc phải lúng túng đề xuất với giám đốc tài chính để xin kinh phí.

Trò chơi nguy hiểm

Ngoài các đột phá về AI, một trong các dự án khiến hội Omega thích thú nhất gần đây là lên kế hoạch kiếm tiền càng nhanh càng tốt sau khi khởi chạy Prometheus. Tất nhiên cơ hội thì đầy rẫy trong nền kinh tế điện tử, nhưng bắt đầu bằng việc làm trò chơi điện tử, âm nhạc, phim ảnh và phần mềm, hay viết sách báo, mua bán chứng khoán, hoặc chế tạo rồi bán các phát minh liệu có tốt hơn không? Mục tiêu của họ về cơ bản là tối đa hóa tỷ lệ lợi tức đầu tư, nhưng các kế hoạch đầu tư thông thường đều là một phiên bản chậm chạp đến mức trào phúng những gì họ đã và đang làm: nếu một nhà đầu tư bình thường có thể hài lòng với lợi tức 9% một năm thì khoản đầu tư MTurk của họ đã sinh ra lợi tức 9% mỗi giờ, lại sinh lãi gấp tám lần mỗi ngày. Giờ họ đã làm bão hòa nguồn cung cấp tại MTurk, vậy tiếp theo sẽ là gì?

Ý nghĩ đầu tiên của họ là đánh quả trên thị trường chứng khoán – dù sao, gần như tất cả bọn họ đều từng từ chối lời mời làm công việc rất lôi cuốn là xây dựng AI cho các quỹ đầu cơ, vốn cũng đang đầu tư mạnh vào chính ý tưởng này. Nếu vài người còn nhớ, đây chính là cách AI kiếm được hàng triệu đô-la đầu tiên trong bộ phim Transcendence (Trí tuệ siêu phàm). Nhưng, các quy định mới về chứng khoán phái sinh⁶ từ sau vụ khủng hoảng năm ngoái⁷ đã hạn chế các lựa chọn của họ. Họ nhanh chóng nhận ra rằng lợi nhuận họ thu được có thể nhiều hơn các nhà đầu tư khác nhưng vẫn khó mà sánh ngang với số tiền thu về từ việc bán sản phẩm của chính mình. Khi có trong tay AI siêu thông minh đầu tiên trên thế giới, ta nên đầu tư vào chính công ty của mình chứ không phải của người khác! Dù đôi lúc ngoại lệ sẽ nảy sinh (như sử dụng khả năng đột nhập siêu phàm của Prometheus để lấy thông tin nội bộ và mua các quyền chọn mua chứng khoán sắp tăng giá), hội Omega cảm thấy việc này không đáng, khi có thể gây ra những chú ý ngoài mong muốn.

Khi họ chuyển hướng tập trung sang các sản phẩm có thể phát triển và bán, trò chơi máy tính dường như là lựa chọn hiển nhiên hàng đầu. Prometheus nhanh chóng trở nên cực kỳ thành thạo trong việc thiết kế các trò chơi hấp dẫn, có thể dễ dàng lập trình, thiết kế đồ họa, thực hiện dò tia hình ảnh cũng như mọi công việc cần thiết khác để tạo ra một sản phẩm hoàn chỉnh, sẵn sàng đưa ra thị trường. Hơn

nữa, sau khi phân tích mọi dữ liệu trên các trang web về sở thích của con người, nó có thể biết chính xác các nhóm người chơi thích gì và phát triển kỹ năng phi thường giúp tối ưu hóa lợi nhuận của một trò chơi. Hồi năm 2011, The Elder Scrolls V: Skyrim (Cổ thư V: Skyrim), trò chơi mà nhiều người trong hội Omega đã chơi nhiều hơn hẳn thời gian họ dám thừa nhận, đã thu về trên 400 triệu đô-la vào tuần đầu tiên. Họ tự tin rằng Prometheus có thể tạo ra trò chơi gây nghiện không kém trong vòng 24 giờ khi sử dụng 1 triệu đô-la tài nguyên trên điện toán đám mây. Họ sẽ bán trò chơi này qua mạng, và Prometheus sẽ đóng giả người chơi để khen ngợi trò chơi này trên cộng đồng mạng. Nếu việc này mang lại 250 triệu đô-la trong một tuần thì xem như họ có thể nhân đôi khoản đầu tư ban đầu tám lần trong tám ngày, đạt lợi tức 3% một giờ – hơi thấp hơn so với khởi đầu ở MTurk, nhưng bền vững hơn nhiều. Bằng việc mỗi ngày phát triển một loạt trò chơi mới, họ dự tính sẽ sớm kiếm được 10 tỷ đô-la, mà không làm bão hòa thị trường trò chơi.

Nhưng một chuyên gia an ninh mạng trong nhóm đã thuyết phục họ từ bỏ trò chơi may rủi này. Cô chỉ ra rằng việc này có thể đem lại rủi ro không thể chấp nhận được là Prometheus sẽ đào tẩu và nắm quyền kiểm soát vận mệnh chính nó. Bởi không thể chắc rằng mục đích của nó sẽ biến đổi thế nào trong chu trình tự cải tiến, họ đã quyết định một lối chơi an toàn và làm mọi cách để giam giữ (hay “đóng hộp”) Prometheus sao cho nó không thể chạy trốn lên Internet. Đối với cỗ máy chính chứa Prometheus trong phòng máy chủ, họ dùng biện pháp cách ly vật lý: nó không có kết nối Internet và đầu ra duy nhất của Prometheus là các tin nhắn và văn bản được nó gửi tới một máy tính do hội Omega điều khiển.

Mặt khác, trên một máy tính có kết nối Internet, chạy bất kỳ chương trình phức tạp nào do Prometheus tạo ra đều có rủi ro: bởi hội Omega không có cách nào hoàn toàn hiểu được nó sẽ làm gì, họ cũng sẽ không biết được nó có làm những việc như tự phát tán nó qua mạng như một loại virus hay không. Khi thử nghiệm phần mềm Prometheus viết cho các công việc trên MTurk, hội Omega đề phòng việc này bằng cách chỉ cho chương trình chạy trên máy ảo. Đây là một chương trình giả lập máy tính: chẳng hạn, rất nhiều người dùng máy Mac sử dụng phần mềm máy ảo để chạy các phần mềm

Windows bằng cách lừa các phần mềm này rằng chúng đang thực sự chạy trên Windows. Hội Omega đã tạo ra máy ảo của chính họ và đặt cho nó biệt danh Hộp Pandora, có chức năng giả lập một máy tính siêu tối giản bỏ đi mọi thiết bị ngoại vi thường thấy: nó không có bàn phím, không màn hình, không loa, không kết nối Internet, không gì cả. Đối với nhiệm vụ chép lại bản thu âm của MTurk, hội Omega thiết lập sao cho thứ duy nhất có thể đi vào Hộp Pandora là một tập tin âm thanh và thứ duy nhất đi ra là một văn bản chữ – chính là bản chép lại. Những giới hạn của hộp này đối với phần mềm trong nó cũng giống như giới hạn của định luật vật lý với chúng ta trong Vũ trụ: phần mềm này không thể di chuyển ra ngoài chiếc hộp, cũng như chúng ta không thể di chuyển nhanh hơn tốc độ ánh sáng, cho dù chúng ta thông minh đến đâu. Ngoại trừ đầu vào và đầu ra duy nhất đó, phần mềm bên trong Hộp Pandora gần như bị đóng chặt vào một vũ trụ song song với các quy tắc điện toán của riêng nó. Hội Omega bị ám ảnh với việc Prometheus đào tẩu đến mức họ còn thêm cả giới hạn thời gian để hạn chế thời gian các mã không tin tưởng hoạt động. Ví dụ, mỗi lần phần mềm chép lại bên trong hộp chép xong một tệp âm thanh, toàn bộ dữ liệu bộ nhớ của Hộp Pandora đều tự động bị xóa và phần mềm được cài đặt lại từ đầu. Như vậy, khi bắt đầu nhiệm vụ chép lại tiếp theo, nó sẽ không biết điều gì đã xảy ra trước đó, do vậy không có khả năng học tập theo thời gian.

Khi hội Omega sử dụng đám mây Amazon cho dự án MTurk, họ có thể đặt mọi mô-đun nhiệm vụ do Prometheus tạo ra vào những chiếc hộp giả lập như vậy trên đám mây, bởi đầu vào và đầu ra của MTurk rất đơn giản. Nhưng họ không thể áp dụng cách này với các trò chơi máy tính đồ họa nặng; chúng không thể bị đóng hộp vì cần có đầy đủ quyền truy cập vào mọi phần cứng máy tính của người chơi. Hơn thế, họ cũng không muốn đánh liều khả năng cho những người dùng máy tính khôn lỏi phân tích mã game của họ, khám phá ra Hộp Pandora và quyết định tìm hiểu thứ bên trong nó. Ở thời điểm này, nguy cơ Prometheus đào tẩu đã loại bỏ không chỉ thị trường trò chơi mà còn cả các thị trường phần mềm lợi nhuận khổng lồ khác, với hàng trăm tỷ đô-la trong tầm tay.

Hàng tỷ đô-la đầu tiên

Hội Omega lúc này đi vào tìm kiếm những sản phẩm giá trị cực cao, dạng kỹ thuật số đơn thuần (để tránh thời gian sản xuất lâu) và dễ nắm bắt (như bài viết hoặc phim ảnh, mà họ biết chắc không gây ra nguy cơ đào tẩu). Cuối cùng, họ quyết định mở một công ty truyền thông, bắt đầu bằng sản phẩm hoạt hình. Trang web, kế hoạch quảng bá và thông cáo báo chí đều đã được chuẩn bị sẵn sàng trước cả khi Prometheus trở nên siêu thông minh – thứ còn thiếu duy nhất là nội dung.

Mặc dù tính tới sáng Chủ Nhật, Prometheus đã có được năng lực đáng kinh ngạc để hốt bạc đều đặn từ MTurk, nhưng khả năng nhận thức của nó vẫn rất hạn chế: người ta vốn chỉ chủ trương tối ưu hóa nó để thiết kế hệ thống AI và viết phần mềm thực hiện các nhiệm vụ MTurk tẻ nhạt. Và như vậy, nó sẽ rất tệ trong khoản làm phim – tệ không phải vì lý do sâu xa nào, mà chính vì lý do James Cameron làm phim tệ khi ông ấy mới sinh ra: đây là một kỹ năng cần được rèn luyện theo thời gian. Giống như một đứa trẻ, Prometheus có thể học bất cứ thứ gì nó muốn từ những dữ liệu truy cập được. Trong khi James Cameron phải mất nhiều năm để học đọc và viết, Prometheus đã lo liệu xong việc này vào Thứ Sáu, trong khi vẫn có thời gian đọc toàn bộ Wikipedia và vài triệu quyển sách. Làm phim lại khó hơn thế. Viết nên một kịch bản mà con người thấy hấp dẫn cũng khó như viết sách vậy, bởi nó đòi hỏi ta phải hiểu biết sâu rộng về xã hội loài người cũng như những yếu tố thú vị đối với con người. Chuyển thể kịch bản thành một tệp video cuối cùng còn đòi hỏi nhiều công sức dò tia hình ảnh các diễn viên mô phỏng và các cảnh phức tạp làm nền cho hoạt động của họ, thêm lời thoại mô phỏng, sản xuất nhạc phim cuốn hút, v.v.. Kể từ sáng Chủ Nhật, Prometheus đã có thể xem hết bộ phim dài hai tiếng chỉ trong một phút, cùng lúc đọc xong cả nguyên tác chuyển thể của bộ phim, tất cả nhận xét và đánh giá trên mạng. Hội Omega nhận thấy sau khi Prometheus liên tục xem vài trăm bộ phim, nó bắt đầu dự đoán khá tốt về việc bộ phim sẽ được đánh giá thế nào và thu hút các đối tượng khán giả khác nhau ra sao.

Nó cũng học được cách tự viết nhận xét phim, mà theo họ, thể hiện sự hiểu biết sâu sắc về mọi thứ từ cốt truyện và diễn xuất tới các chi tiết kỹ thuật như ánh sáng và góc quay. Họ cho rằng việc này sẽ đảm bảo khi Prometheus làm phim, nó sẽ biết làm thế nào để thành công.

Hội Omega chỉ đạo Prometheus tập trung làm phim hoạt hình trước, để tránh những câu hỏi khó xử rằng diễn viên mô phỏng là ai. Vào tối Chủ Nhật, họ chốt lại giai đoạn làm việc cuối tuần cuồn cuộn đại với bia và bỏng ngô từ lò vi sóng, tắt bớt đèn để xem bộ phim đầu tay của Prometheus. Đó là một bộ phim hoạt hình hài kịch kỳ ảo tương tự như Frozen (Nữ hoàng băng giá) của Disney, được phần mềm do Prometheus viết đảm nhận việc dò tia, chạy trên môi trường đóng hộp của đám mây Amazon; việc này tiêu tốn gần hết 1 triệu đô-la lợi nhuận MTurk của ngày hôm đó. Khi bộ phim bắt đầu, họ thấy vừa thú vị vừa lo sợ vì sắp được xem bộ phim do một cỗ máy không cần con người chỉ đạo sản xuất. Tuy vậy, họ nhanh chóng bật cười trước những màn tấu hài và nín thở trước những cảnh gay cấn. Một vài người còn rơm rớm nước mắt ở phân cảnh kết thúc cảm động; họ như bị cuốn hoàn toàn vào hiện thực hư cấu, đến mức quên mất thứ gì đã sản xuất bộ phim này.

Hội Omega lên kế hoạch mở cửa trang web vào Thứ Sáu, để có thời gian cho Prometheus sản xuất thêm nội dung và để bản thân họ rảnh rang làm những việc không tin tưởng giao cho Prometheus: mua quảng cáo và tuyển dụng nhân viên cho công ty vỏ bọc thành lập từ nhiều tháng trước. Để che dấu, họ đưa ra hoàn cảnh chính thức là: công ty truyền thông của họ (không hề có liên quan công khai nào với hội Omega) mua phần lớn nội dung từ các nhà sản xuất phim độc lập, chủ yếu là các công ty khởi nghiệp công nghệ cao ở các khu vực thu nhập thấp. Cũng tiện lắm thay, các nguồn cung giả này đều nằm ở các vùng hẻo lánh như Tiruchchirappalli và Yakutsk, nơi mà những phóng viên tò mò nhất cũng không buồn tới thăm. Những nhân viên thực sự được tuyển sẽ chịu trách nhiệm về quảng cáo và hành chính; họ sẽ trả lời tất cả mọi người rằng đội sản xuất ở một địa điểm khác và không nhận phỏng vấn vào thời điểm này. Để phù hợp với hoàn cảnh đó, họ chọn khẩu hiệu công ty là “Lan tỏa tài năng sáng tạo thế giới”, tạo dựng hình ảnh công ty rất tân thời và

khác biệt qua việc ứng dụng công nghệ tiên tiến nhằm hỗ trợ những người sáng tạo, đặc biệt là ở các nước đang phát triển.

Khi Thứ Sáu tới, những vị khách tò mò bắt đầu truy cập vào địa chỉ và bất ngờ trước một trang web tương tự dịch vụ giải trí qua mạng như Netflix và Hulu nhưng lại có đầy những khác biệt thú vị. Tất cả các bộ phim hoạt hình dài tập đều là những bộ mới họ chưa từng biết đến. Chúng đều khá hấp dẫn: hầu hết các tập trong mỗi bộ đều dài 45 phút với cốt truyện chặt chẽ, mỗi tập đều có kết cục khiến bạn háo hức muốn tìm hiểu diễn biến tập tiếp theo. Và chúng có giá rẻ hơn các hãng đối thủ. Tập đầu tiên của mỗi bộ phim thì miễn phí, những tập khác có giá 49 xu một tập, có khuyến mại khi thuê cả bộ. Ban đầu trang web chỉ có ba bộ phim, mỗi bộ ba tập, nhưng các tập mới được thêm vào hàng ngày, cùng các bộ phim mới để phục vụ nhiều đối tượng khác nhau. Trong hai tuần đầu tiên, kỹ năng làm phim của Prometheus cải thiện nhanh chóng, không chỉ về chất lượng phim mà cả về thuật toán để giả lập nhân vật và dò tia, nhờ đó giảm thiểu chi phí cho điện toán đám mây để sản xuất mỗi tập. Vì lẽ đó, hội Omega có thể phát hành hàng chục bộ phim mới trong tháng đầu tiên, hướng tới mọi đối tượng từ trẻ em tới người lớn, cũng như mở rộng tới mọi thị trường ngôn ngữ lớn trên thế giới, giúp trang web của họ có tính quốc tế hơn tất cả các đối thủ. Một số nhà bình luận bị ấn tượng bởi các bộ phim không chỉ được lồng tiếng nhiều ngôn ngữ mà ngay cả phần hình ảnh cũng vậy: ví dụ, khi một nhân vật nói tiếng Ý, cử động miệng sẽ khớp với từ ngữ và ngôn ngữ cơ thể điển hình của người Ý. Mặc dù Prometheus hiện tại hoàn toàn có năng lực làm các bộ phim với diễn viên mô phỏng giống hệt con người, hội Omega vẫn né tránh việc này để khỏi vô tình làm lộ bí mật. Tuy nhiên, họ cũng cho ra đời nhiều bộ phim với các nhân vật hoạt hình bán thực, ở các thể loại cạnh tranh với các chương trình tivi và phim ảnh truyền thống do người đóng.

Kênh phim của họ hóa ra khá thu hút và đạt mức tăng trưởng người xem ngoạn mục. Nhiều người hâm mộ cảm thấy nhân vật và cốt truyện ở đó thông minh và thú vị hơn cả những tác phẩm Hollywood màn ảnh rộng đắt đỏ. Họ cũng rất hân hoan, vì có thể xem phim với giá cả phải chăng hơn nhiều. Nổi lên nhờ quảng cáo rầm rộ (hội Omega có khả năng chi trả khoản này do phí tổn sản xuất gần

như bằng không), truyền thông xuất sắc và nhận xét truyền miệng mạnh mẽ, doanh thu toàn cầu của họ đã tăng tốc tới mức 10 triệu đô-la một ngày chỉ sau một tháng hoạt động. Sau hai tháng, họ đã vượt qua Netflix, và sau ba tháng, họ đã thu về hơn 100 triệu đô-la một ngày, trở thành một trong những đế quốc truyền thông lớn nhất thế giới, bắt đầu cạnh tranh với Time Warner, Disney, Comcast và Fox.

Thành công gây xôn xao dư luận của họ đã thu hút nhiều sự chú ý không mong muốn, bao gồm cả suy đoán về việc họ có công cụ AI mạnh. Nhưng chỉ cần dùng một phần nhỏ doanh thu, hội Omega đã phát động một chiến dịch đánh lạc hướng khá thành công. Từ một văn phòng hào nhoáng mới ở Manhattan, các phát ngôn viên mới tuyển của họ thêm mắm dặm muối cho câu chuyện che đậy. Hàng loạt người được tuyển để làm nền; kể cả những biên kịch thực sự trên khắp thế giới cũng được thuê để xây dựng các bộ phim mới. Tất cả đều không biết gì về Prometheus. Mạng lưới thầu phụ quốc tế phức tạp cũng dễ dàng khiến phần lớn các nhân viên tin rằng có người ở đâu đó đang làm phần lớn mọi việc.

Để khỏi bị lộ và bị nghi ngờ do sử dụng điện toán đám mây quá mức, họ còn thuê các kỹ sư xây dựng một chuỗi các cơ sở máy tính khổng lồ quanh thế giới, do các công ty vỏ bọc tưởng như không có mối liên kết nào sở hữu. Mặc dù được quảng cáo với dân địa phương là “các trung tâm dữ liệu xanh” vì phần lớn chúng tiêu thụ năng lượng mặt trời, nhưng các cơ sở này thực ra tập trung chủ yếu vào việc tính toán hơn là lưu trữ. Prometheus đã thiết kế từng chi tiết nhỏ nhất cho những trung tâm này, chỉ sử dụng các phần cứng có sẵn và tối ưu hóa chúng để giảm thiểu thời gian xây dựng. Những người xây dựng và quản lý các trung tâm này không hề biết thứ gì được tính toán ở đây: họ nghĩ mình chỉ đang quản lý các cơ sở điện toán đám mây thương mại tương tự như của Amazon, Google hay Microsoft và chỉ biết rằng việc kinh doanh được quản lý từ xa.

Công nghệ mới

Trong vài tháng, đế chế kinh doanh kiểm soát bởi hội Omega bắt đầu có địa vị chắc chắn ở nhiều lĩnh vực kinh tế quốc tế, nhờ khả năng lên kế hoạch siêu phàm của Prometheus. Bằng cách phân tích kỹ lưỡng dữ liệu trên thế giới, từ tuần đầu tiên, nó đã cung cấp cho hội Omega một kế hoạch phát triển từng bước và tiếp tục cải thiện, trau chuốt kế hoạch này mỗi khi dữ liệu cũng như tài nguyên máy tính tăng lên. Mặc dù Prometheus còn chưa tới mức thông tuệ mọi thứ nhưng khả năng của nó bây giờ đã vượt xa con người, tới mức hội Omega coi nó như nhà tiên tri hoàn hảo, luôn cần mẫn đưa ra những lời giải đáp và lời khuyên sáng suốt cho mọi câu hỏi.

Phần mềm của Prometheus hiện tại đã được tối ưu hóa cao để tận dụng tối đa phần cứng tầm thường do con người phát minh ra cho nó, và như hội Omega dự đoán, Prometheus đã tìm ra nhiều cách cải thiện phần cứng này. Do lo sợ việc nó vượt khỏi tầm kiểm soát, họ không xây dựng các cơ sở người máy để Prometheus điều hành trực tiếp. Thay vào đó, nhóm thuê nhiều nhà khoa học và kỹ sư hàng đầu ở nhiều nơi trên thế giới, cung cấp cho họ các báo cáo nghiên cứu nội bộ mà Prometheus viết ra, vờ rằng chúng là của các nhà nghiên cứu ở các bộ phận khác. Những báo cáo này viết rất rõ các tác động vật lý và kỹ thuật sản xuất mới mẻ – các công nghệ này sẽ được các kỹ sư nhanh chóng thử nghiệm, hiểu rõ và làm chủ. Dĩ nhiên, quy trình nghiên cứu và phát triển (R&D) của con người thường tốn hàng năm trời, bởi phải qua nhiều vòng thử-sai chậm chạp. Nhưng tình hình hiện tại khác hẳn: Prometheus đã biết trước bước tiếp theo, vì vậy yếu tố cản trở chỉ đơn giản là sau khi được hướng dẫn, con người sẽ mất bao lâu để hiểu và xây dựng đúng thứ. Một giáo viên giỏi có thể giúp học sinh học khoa học tiến bộ nhanh hơn so với tự mình khám phá từ con số không, và Prometheus đã bí mật làm điều tương tự với các nhà nghiên cứu. Vì Prometheus có thể dự đoán chính xác khoảng thời gian con người cần để hiểu và xây dựng mọi thứ nếu có công cụ nên nó đã xác lập con đường nhanh nhất có thể, ưu tiên những công cụ vừa dễ dàng nắm bắt và xây dựng, vừa có ích cho việc phát triển các công cụ cao cấp hơn.

Hưởng ứng phong trào sáng chế, các đội ngũ kỹ sư được khuyến khích sử dụng chính máy móc của họ để xây dựng những máy móc tốt hơn. Việc tự cung cấp không chỉ tiết kiệm tiền bạc mà còn giúp

tránh được nguy cơ từ thế giới bên ngoài trong tương lai. Trong vòng hai năm, họ đã sản xuất ra các phần cứng máy tính tốt hơn hẳn những gì thế giới từng biết đến. Để tránh tiếp tay cho các đối thủ bên ngoài, họ giữ kín công nghệ này và chỉ dùng nó để nâng cấp Prometheus.

Mặc dù vậy, thế giới cũng nhận ra rằng một cuộc bùng nổ công nghệ đáng kinh ngạc đang diễn ra. Các công ty mới nổi trên khắp thế giới đang tung ra nhiều sản phẩm cách mạng ở gần như mọi lĩnh vực. Một công ty khởi nghiệp của Hàn Quốc giới thiệu loại pin mới, với mức trữ năng lượng gấp bốn lần pin máy tính trên thị trường nhưng có thể sạc đầy trong vòng chưa tới một phút. Một hãng Phần Lan ra mắt loại pin quang điện với hiệu suất chuyển đổi gấp đôi đối thủ mạnh nhất. Một công ty Đức công bố loại dây siêu dẫn ở nhiệt độ phòng có thể sản xuất hàng loạt, tạo nên cuộc cách mạng trong ngành năng lượng. Một tập đoàn công nghệ sinh học ở Boston thông báo thử nghiệm lâm sàng giai đoạn II loại thuốc giảm cân mà theo họ là loại đầu tiên thực sự hiệu quả và không gây tác dụng phụ, trong khi theo tin đồn thì một đơn vị Ấn Độ cũng đang bán loại thuốc tương tự ở chợ đen. Một công ty California cũng phản pháo bằng việc thử nghiệm giai đoạn II cho một loại thuốc chống ung thư đột phá, có thể giúp hệ miễn dịch của cơ thể nhận biết và tấn công bất kỳ tế bào nào có đột biến gene ung thư phổ biến. Các ví dụ còn tiếp nối, dẫn tới đồn đoán về một kỷ nguyên vàng mới trong ngành khoa học. Sau cùng nhưng không kém phần quan trọng, các công ty robot mọc lên như nấm trên khắp thế giới. Không người máy nào đạt tới gần trí tuệ loài người và đa phần đều trông không giống con người. Nhưng, chúng đã đột ngột gây rối loạn nền kinh tế, và trong nhiều năm tiếp theo, chúng dần dần thay thế hầu hết công nhân trong các ngành sản xuất, vận tải, lưu kho, bán lẻ, xây dựng, khai thác và nông – lâm – ngư nghiệp.

Nhờ công lao của một đội ngũ luật sư xuất sắc, thế giới đã không nhận ra là tất cả các công ty này đều được hội Omega điều khiển thông qua một chuỗi trung gian. Prometheus đã khiến các văn phòng sáng chế trên thế giới tràn ngập những phát minh gây chấn động được gửi từ rất nhiều người đại diện và những phát minh này từ từ thống trị mọi lĩnh vực công nghệ.

Các công ty gây biến động thị trường này không chỉ gây thù chuốc oán với nhiều đối thủ cạnh tranh mạnh mẽ, mà chúng còn kết thân được với những người bạn mạnh hơn. Các công ty này có lợi nhuận rất cao và với khẩu hiệu kiểu như “Đầu tư vào cộng đồng chúng ta”, họ dành ra một phần doanh thu đáng kể để tuyển người thực hiện các dự án cộng đồng – thường lại chính là những người bị sa thải từ các công ty ảnh hưởng bởi biến động. Họ sử dụng phân tích chi tiết của Prometheus để giao những công việc bổ ích nhất cho các nhân sự và cộng đồng với chi phí thấp nhất, phù hợp với hoàn cảnh địa phương. Tức là những khu vực có tỷ trọng dịch vụ công cao, công việc được giao sẽ tập trung vào xây dựng cộng đồng, văn hóa và chăm sóc người bệnh, trong khi ở khu vực nghèo hơn thì công việc sẽ bao gồm cả xây dựng và duy trì trường học, dịch vụ sức khỏe, chăm sóc trẻ em, chăm sóc người già, nhà giá rẻ, công viên và cơ sở hạ tầng cơ bản. Gần như ở mọi nơi, người dân địa phương đồng tình rằng việc này lẽ ra phải được thực hiện từ lâu. Các chính khách địa phương thì nhận được những khoản quyên góp lớn và mọi việc đều được lo liệu để những người này luôn có được hình ảnh tốt đẹp từ việc khuyến khích các tập đoàn đầu tư vào cộng đồng.

Vươn tới quyền lực

Hội Omega thành lập công ty truyền thông không chỉ để hỗ trợ tài chính cho các dự án công nghệ mới của họ, mà còn để chuẩn bị cho bước tiếp theo của kế hoạch táo bạo: tiếp quản thế giới. Trong vòng một năm sau khi bắt đầu, họ đã thêm nhiều kênh tin tức vô cùng đặc sắc vào mạng lưới rộng khắp của mình. Đối lập với các kênh khác của họ, những kênh này được thiết kế với mục đích tiêu tiền và được gán cho cái mác dịch vụ công. Trên thực tế, các kênh tin tức của nhóm không đem lại bất kỳ nguồn thu nào: không có quảng cáo và bất cứ ai có kết nối mạng đều được xem miễn phí. Toàn bộ phần còn lại của đế chế truyền thông Omega là cỗ máy kiếm tiền mạnh tới mức họ có thể tiêu tốn nhiều tài nguyên cho dịch vụ tin tức hơn bất cứ báo chí nào trong lịch sử – và điều đó được thể hiện rõ. Bằng việc

ráo rức tuyển dụng và kê ra mức lương cạnh tranh cao cho các ký giả và phóng viên điều tra, họ đã đưa được lên màn ảnh nhiều tài năng và phát hiện nổi bật. Nhờ một dịch vụ web quốc tế trả tiền cho bất cứ ai cung cấp tin tức đáng chú ý, họ luôn là người đầu tiên đưa tin, từ các vụ tham nhũng ở địa phương tới những sự kiện gây xúc động. Ít nhất đó là những gì mọi người tin vào, còn trên thực tế, họ là người đầu tiên vì những câu chuyện vốn được cho là của các ký giả công dân đều được chính Prometheus khám phá thông qua giám sát thường trực trên Internet. Tất cả các trang tin tức video đều có kênh podcast⁸ và các bài viết.

Giai đoạn 1 của chiến thuật tin tức là đạt được lòng tin của mọi người, và họ đã rất thành công. Tâm lý sẵn sàng đổ nhiều tiền chưa từng có của họ đã khiến lượng tin bài địa phương và khu vực luôn dày đặc đến khó tin, với các bài điều tra thường phơi bày những vụ bê bối thực sự thu hút người xem. Bất cứ khi nào một quốc gia bị chia cắt sâu sắc về mặt chính trị và quen với thông tin đảng phái, họ sẽ mở một kênh tin tức dành riêng cho từng phe, nhìn bề ngoài thì mỗi kênh lại thuộc sở hữu của một công ty khác nhau, rồi dần dần thu phục lòng tin của từng đảng phái. Nếu có thể, họ sẽ làm việc này thông qua người đại diện, mua lại các kênh đang hoạt động có tầm ảnh hưởng nhất và dần dần cải thiện chúng bằng việc loại bỏ quảng cáo và giới thiệu các nội dung riêng. Còn ở những quốc gia có công tác kiểm duyệt và can thiệp chính trị đe dọa những nỗ lực này, họ thường bước đầu đồng ý với bất kỳ yêu cầu nào của chính phủ để tồn tại, tuân theo câu khẩu hiệu nội bộ bí mật “Sự thật, không gì ngoài sự thật, nhưng có thể không phải toàn bộ sự thật”. Prometheus thường đưa ra những lời khuyên tuyệt vời trong trường hợp này, chỉ rõ những chính trị gia nào cần được tô vẽ và ai (thường là những kẻ tham ô ở địa phương) cần phải vạch mặt. Nó cũng cung cấp những đề nghị giá trị cho việc nên giật dây việc gì, hối lộ ai và làm điều đó thế nào cho hiệu quả nhất.

Chiến thuật này thành công rực rỡ trên khắp thế giới, khi mà các kênh thuộc kiểm soát của hội Omega dần được coi là nguồn tin tức đáng tin cậy nhất. Cũng có những chính phủ cản trở để các kênh này không thể phổ biến rộng rãi; nhưng ngay cả ở quốc gia đó, chúng

vẫn xây dựng được lòng tin với công chúng và có rất nhiều phóng sự lan tỏa trên mọi con phố. Giám đốc các hãng tin tức đối thủ đều cảm thấy họ đang tham gia một cuộc chiến vô vọng: sinh lời bằng cách nào khi phải đối chọi với một kẻ đã có tiềm lực tốt lại còn đem sản phẩm tặng không? Bởi lượng người xem sứt giảm nên ngày càng nhiều mạng lưới quyết định bán kênh tin tức – thường là cho các liên doanh mà hóa ra vẫn do hội Omega quản lý.

Sau khoảng hai năm Prometheus đi vào hoạt động và phần lớn giai đoạn chiếm được lòng tin đã hoàn tất, hội Omega tiến hành giai đoạn 2 trong chiến thuật tin tức: thuyết phục. Từ trước đó, nhiều người quan sát sắc sảo đã nhận thấy dấu hiệu của một ý đồ chính trị đằng sau các kênh tin tức: dường như có một lực đẩy nhẹ nhàng về hướng trung tâm, rời xa mọi phương hướng cực đoan. Mạng lưới kênh tin tức dành cho các đối tượng khác nhau của họ vẫn phản ánh tình trạng thù địch giữa Mỹ và Nga, Ấn Độ và Pakistan, giữa các tôn giáo, các đảng phái chính trị, v.v., nhưng giọng điệu chỉ trích đã nhẹ nhàng hơn và thường tập trung vào các vấn đề cụ thể như tiền bạc và quyền lực chứ không còn là các công kích cá nhân hay tin đồn gây hoang mang và thiếu căn cứ nữa. Ngay khi giai đoạn 2 bắt đầu, lực đẩy nhằm giảm thiểu những xung đột cũ này càng trở nên rõ ràng hơn qua các câu chuyện cảm động về cảnh ngộ của những kẻ thù truyền kiếp đi kèm các phóng sự điều tra về việc có bao nhiêu kẻ gây xung đột chỉ vì động cơ lợi nhuận cá nhân.

Các nhà bình luận chính trị nhận thấy rằng, bên cạnh việc xoa dịu các xung đột khu vực, dường như có một lực đẩy được tính toán kỹ lưỡng nhằm giảm thiểu những mối đe dọa toàn cầu. Ví như, rủi ro từ chiến tranh hạt nhân bỗng nhiên được đưa ra thảo luận ở khắp nơi. Vài bộ phim bom tấn lấy bối cảnh chiến tranh hạt nhân toàn cầu xảy ra do sơ suất hoặc có chủ đích và xoáy vào hậu quả thảm khốc bao gồm mùa đông hạt nhân, cơ sở hạ tầng sụp đổ và chết đói hàng loạt. Những bộ phim tài liệu mới đây thuyết phục lại trình bày tỉ mỉ việc mùa đông hạt nhân có thể tác động đến mọi quốc gia như thế nào. Những nhà khoa học và chính trị gia ủng hộ giảm leo thang hạt nhân được đưa lên thời sự liên tục, chủ yếu để thảo luận kết quả của các nghiên cứu mới về những biện pháp hữu ích có thể áp dụng – chính các tổ chức khoa học nhận tài trợ lớn từ các công ty công nghệ mới

đã tài trợ thực hiện những nghiên cứu này. Kết quả là, động lực chính trị nhằm tháo bỏ tên lửa khỏi bộ phóng kích hoạt nhạy qua báo động⁹ và thu hẹp kho vũ khí hạt nhân bắt đầu lớn mạnh. Các phương tiện truyền thông cũng chuyển hướng tập trung sang tình hình biến đổi khí hậu toàn cầu, thường theo cách nhấn mạnh vào những đột phá công nghệ mới nhất từ Prometheus đã giảm chi phí năng lượng tái tạo và khuyến khích các chính phủ đầu tư vào kiểu cơ sở hạ tầng năng lượng mới này.

Đồng thời với việc tiếp quản truyền thông, hội Omega còn tận dụng Prometheus để cách mạng hóa giáo dục. Nếu biết kiến thức và khả năng của một người, Prometheus sẽ xác định được cách nhanh nhất giúp họ học bất kỳ môn học mới nào, với tinh thần quyết tâm và niềm thích thú mà tiếp tục, và đưa ra các video, tài liệu đọc, bài tập và các công cụ học tập được tối ưu hóa cho đối tượng. Vì lẽ đó, các công ty do nhóm Omega kiểm soát có thể đưa ra các khóa học trực tuyến về hầu hết chủ đề, được tùy biến cao cả về ngôn ngữ, văn hóa lẫn trình độ bắt đầu. Dù bạn là một người 40 tuổi mù chữ muốn học đọc, hay một tiến sĩ sinh học đang tìm kiếm thông tin mới nhất về liệu pháp miễn dịch trong điều trị ung thư, Prometheus đều có khóa học hoàn hảo cho bạn. Các dịch vụ này khác hẳn những khóa học trực tuyến hiện đại nhất hiện nay: tài làm phim của Prometheus được tận dụng để tạo nên các phân đoạn video thực sự cuốn hút, chứa đựng những ẩn dụ mạnh mẽ giúp bạn dễ liên tưởng và khiến bạn khao khát học tiếp. Một số khóa học được bán lấy lợi nhuận, nhưng nhiều khóa được cung cấp miễn phí; điều này khiến rất nhiều người hân hoan, từ các giáo viên trên toàn thế giới sẽ sử dụng chúng trong lớp học của họ tới hầu hết những người hăng hái học tập.

Siêu năng lực trong giáo dục này tỏ rõ là công cụ mạnh mẽ để phục vụ mục đích chính trị, khi tạo ra “chuỗi thuyết phục” trực tuyến: nội dung chi tiết của mỗi video sẽ vừa đổi mới cách nhìn của người xem, vừa khuyến khích họ xem một video khác với chủ đề liên quan có khả năng thuyết phục họ hơn nữa. Giả dụ, nếu cần giải quyết xung đột giữa hai quốc gia, các phim tài liệu lịch sử sẽ được phát hành độc lập ở cả hai nước, trong đó người ta sẽ nhận xét về nguồn gốc và diễn biến xung đột đó từ nhiều sắc thái hơn. Những phóng sự

mang tính giáo dục sẽ giải thích ai ở chính phe của họ thực sự hưởng lợi từ xung đột liên tục và phương pháp mà người đó kích động xung đột xảy ra. Đồng thời, các nhân vật đáng mến ở nước bên kia sẽ bắt đầu xuất hiện trong những chương trình nổi tiếng của các kênh giải trí, đúng theo cách các nhân vật ở phe thiểu số với hình ảnh thiện cảm đã giúp ủng hộ các phong trào dân quyền và quyền đồng tính trong quá khứ.

Chẳng bao lâu, các nhà bình luận chính trị sẽ không thể bỏ qua sự ủng hộ ngày càng tăng dành cho một chương trình nghị sự chính trị tập trung vào bảy khẩu hiệu:

1. Dân chủ
2. Cắt giảm thuế
3. Cắt giảm dịch vụ xã hội của chính phủ
4. Cắt giảm chi tiêu quân sự
5. Tự do thương mại
6. Mở cửa biên giới
7. Các công ty có trách nhiệm xã hội

Sẽ khó mà nhận ra mục đích đằng sau những khẩu hiệu này lại là làm xói mòn tất cả các cấu trúc quyền lực trước nay trên thế giới. Các mục 2 – 6 làm xói mòn quyền lực nhà nước và việc dân chủ hóa thế giới đã cho đế chế kinh doanh của hội Omega nhiều lợi thế hơn trong việc lựa chọn các lãnh đạo chính trị. Các công ty có trách nhiệm xã hội làm suy yếu quyền lực nhà nước sâu hơn bằng cách tiếp quản ngày càng nhiều các dịch vụ mà chính phủ vốn (phải) cung cấp. Tầng lớp tinh hoa kinh doanh kiểu truyền thống sẽ bị suy yếu, bởi không có khả năng cạnh tranh với các công ty được Prometheus hậu thuẫn trên thị trường tự do, do đó sẽ chiếm tỷ trọng giảm dần trong tổng kinh tế thế giới. Các nhà lãnh đạo mang quan điểm truyền thống, từ các đảng phái chính trị cho đến các nhóm tín ngưỡng, thì lại không có công cụ thuyết phục nào cạnh tranh nổi với đế chế truyền thông của hội Omega.

Trong mọi thay đổi sâu rộng đều có người thắng và kẻ thua. Một tinh thần lạc quan mới đã xuất hiện rõ ràng ở hầu hết các quốc gia, khi giáo dục, dịch vụ xã hội, cơ sở hạ tầng được cải thiện, các cuộc xung đột giảm xuống và các công ty địa phương tung ra những công nghệ đột phá lan tỏa khắp thế giới. Tuy vậy, không phải ai cũng vui

về. Trong khi nhiều công nhân bị mất việc được thuê làm lại cho các dự án cộng đồng thì những người nắm giữ quyền lực và tiền bạc lại nhận thấy cả hai thứ này dần thu hẹp. Hiện tượng này khởi đầu từ lĩnh vực truyền thông và công nghệ, rồi lan rộng ra hầu hết các lĩnh vực khác. Xung đột trên thế giới suy giảm dẫn tới việc cắt giảm ngân sách quốc phòng, gây tổn hại cho các nhà thầu quân sự. Các công ty mới nổi đang phát triển thường không bán cổ phần ra đại chúng, với lý do các cổ đông chạy theo lợi nhuận sẽ ngăn cản họ đổ tiền ào ạt vào các dự án cộng đồng. Vì vậy, thị trường chứng khoán toàn cầu liên tục mất giá, đe dọa cả các ông trùm tài chính lẫn những người dân bình thường dựa dẫm vào quỹ lương hưu. Như thể lợi nhuận sụt giảm của các công ty cổ phần đại chúng là chưa đủ tồi tệ, các công ty đầu tư trên toàn thế giới còn nhận thấy một xu hướng đáng lo ngại: tất cả các thuật toán giao dịch thành công trước đó dường như đã ngừng hoạt động, hay hoạt động kém hiệu quả ngay cả ở quỹ chỉ số đơn giản. Dường như luôn có một kẻ nào đó cao tay hơn họ và đánh bại họ ngay tại sân nhà.

Cho dù đa số những người quyền lực chống lại làn sóng thay đổi này, phản ứng của họ thực sự chẳng thấm thía gì, như thể họ đã rơi vào một cái bẫy được lên kế hoạch chu đáo. Những thay đổi lớn diễn ra với tốc độ đáng kinh ngạc đến nỗi người ta khó lòng theo kịp để phối hợp xây dựng biện pháp đối phó. Hơn nữa, họ không rõ nên phát cò để đấu tranh cho điều gì. Phe chính trị cánh hữu truyền thống nhận thấy phần lớn khẩu hiệu của họ được đồng thuận, nhưng việc cắt giảm thuế và cải thiện môi trường kinh doanh lại chủ yếu giúp ích cho đối thủ cạnh tranh của họ, những người có công nghệ cao hơn. Hầu như tất cả các ngành công nghiệp truyền thống đều kêu gọi cứu trợ, nhưng ngân quỹ eo hẹp của chính phủ chỉ khiến họ thêm đáng thương bởi phải tranh giành lẫn nhau đầy vô vọng, trong khi phương tiện truyền thông thì miêu tả họ như những con khủng long tìm kiếm trợ cấp của nhà nước vì không còn cạnh tranh nổi. Phe chính trị cánh tả truyền thống phản đối việc tự do thương mại và cắt giảm dịch vụ xã hội của chính phủ, nhưng vui mừng trước tình hình cắt giảm quân sự và giảm thiểu đói nghèo. Thực chất, phần lớn sự tức giận trong họ đều đã bị lấn lướt bởi thực tế không thể phủ nhận là các dịch vụ xã hội hiện nay cải thiện hơn khi chúng được các công

ty lý tưởng cung cấp chứ không phải nhà nước. Hết thăm dò ý kiến này tới khảo sát khác đều cho thấy hầu hết các cử tri trên khắp thế giới cảm thấy chất lượng cuộc sống của họ được cải thiện và nhìn chung mọi thứ đang đi theo chiều hướng tốt. Một phép tính đơn giản có thể giải thích việc này: trước khi có Prometheus, 50% dân số Trái đất nghèo nhất chỉ kiếm được khoảng 4% thu nhập toàn cầu, vì thế các công ty do hội Omega kiểm soát chỉ cần chia sẻ một phần rất nhỏ lợi nhuận với nhóm này là có thể giành được trái tim của họ (và cả phiếu bầu).

Hợp nhất

Kết quả là, từ quốc gia này đến quốc gia khác các đảng phái đi theo bầy khẩu hiệu của hội Omega đều chiến thắng ngoạn mục ở các cuộc bầu cử. Trong các chiến dịch được tối ưu kỹ càng, họ xây dựng hình tượng ở vị trí trung tâm của phổ chính trị, tố cáo cánh phải là những kẻ thích gây hấn tham lam chuyên đòi hỏi cứu trợ và lên án cánh trái là những tên ngăn chặn đổi mới chỉ thích tăng thuế và chi tiêu cho hoạt động chính phủ. Hầu như không ai biết rằng Prometheus đã cẩn thận lựa chọn những người có ưu thế để chuẩn bị cho vị trí ứng cử viên và sử dụng mọi ảnh hưởng để đảm bảo họ chiến thắng.

Trước thời Prometheus, ngày càng nhiều người ủng hộ phong trào phổ cập thu nhập cơ bản để khắc phục tình trạng thất nghiệp do công nghệ. Họ đề xuất cấp cho mọi người một khoản thu nhập tối thiểu trích từ thuế. Phong trào này đổ sụp khi dự án cộng đồng của các công ty khởi động, vì đế chế kinh doanh do hội Omega kiểm soát đang làm chính điều tương tự. Với lý do tăng cường sự phối hợp giữa các dự án cộng đồng, một nhóm công ty quốc tế đã thành lập Liên minh Nhân đạo, một tổ chức phi chính phủ có mục tiêu xác minh và tài trợ cho những chương trình nhân đạo có giá trị nhất trên toàn thế giới. Chẳng bao lâu, hầu như toàn bộ đế chế Omega đều ủng hộ Liên minh này; họ tung ra các dự án toàn cầu với quy mô chưa từng có, ở ngay cả những nước đã lỡ mất phần lớn cơ hội trong cuộc

bùng nổ công nghệ, nhằm cải thiện giáo dục, y tế, quản trị và thúc đẩy thịnh vượng. Dĩ nhiên, chính Prometheus đã cẩn thận xây dựng các kế hoạch dự án đằng sau hậu trường, xếp hạng chúng theo tác động tích cực trên mỗi đô-la. Thay vì chỉ đơn giản là đưa tiền mặt như đề xuất trong chương trình thu nhập cơ bản, Liên minh (theo cách gọi thân thiện) thu hút những người được hỗ trợ đến làm việc cho mục tiêu của họ. Kết quả là, phần lớn dân số thế giới cảm thấy biết ơn và trung thành với Liên minh – còn hơn cả chính phủ của họ.

Theo thời gian, Liên minh dần dần đảm nhận vai trò của một chính phủ thế giới, trong khi chính phủ các quốc gia đều nhận thấy quyền lực của họ liên tục suy giảm. Ngân sách quốc gia tiếp tục thu hẹp do cắt giảm thuế, trong khi ngân quỹ của Liên minh tăng trưởng tới mức vượt cả tổng ngân sách của tất cả các chính phủ. Mọi vai trò truyền thống của chính phủ quốc gia ngày càng trở nên dư thừa và không còn thích hợp. Liên minh đã cung cấp dịch vụ xã hội, giáo dục và cơ sở hạ tầng tốt nhất từ trước đến nay. Truyền thông đã xoa dịu xung đột quốc tế đến mức phần lớn chi phí quân sự trở nên không cần thiết, và sự thịnh vượng gia tăng đã nhổ bỏ phần lớn gốc rễ các xung đột cũ, vốn xuất phát từ việc cạnh tranh tài nguyên khan hiếm. Một số tên độc tài cùng những kẻ khác lại phản đối mãnh liệt trật tự thế giới mới này và từ chối bị mua chuộc, nhưng tất cả đều bị lật đổ trong những cuộc đảo chính được dàn dựng cẩn thận hoặc những cuộc nổi dậy hàng loạt.

Hội Omega lúc này đã hoàn thành quá trình chuyển đổi ấn tượng nhất trong lịch sử cuộc sống trên Trái Đất. Lần đầu tiên từ trước tới nay, hành tinh của chúng ta được một thế lực duy nhất điều hành, được khuếch đại sức mạnh bởi một trí tuệ cao siêu đến nỗi nó có khả năng giúp cuộc sống phát triển hàng tỷ năm trên Trái Đất và xuyên suốt vũ trụ –nhưng kế hoạch cụ thể của họ là gì?

—ooOoo—

Đó là câu chuyện của nhóm Omega. Phần còn lại của cuốn sách này nói về một câu chuyện khác – một câu chuyện hãy còn chưa được viết: tương lai của chính chúng ta với AI. Bạn muốn nó diễn ra như thế nào? Liệu điều xa lạ như câu chuyện của nhóm Omega có

khả năng xảy ra không, và nếu có thì bạn muốn nó xảy ra chứ? Bỏ qua những phán đoán về AI siêu phàm, bạn muốn câu chuyện của chúng ta bắt đầu như thế nào? Bạn muốn AI tác động đến công việc, luật pháp và vũ khí trong thập kỷ tới theo cách nào? Nhìn xa hơn về phía trước, bạn sẽ viết nên kết cục như thế nào? Đây sẽ là một câu chuyện thực sự vĩ đại, bởi nó không chứa đựng điều gì ngoài chính vận mệnh tương lai của sự sống trong vũ trụ chúng ta. Và đó là câu chuyện chúng ta phải viết.

Chương 1

Chào mừng tới cuộc đối thoại quan trọng nhất của thời đại

Công nghệ mang lại cho sự sống tiềm năng để phát triển chưa từng có – hoặc để tự hủy diệt.

Viện Tương lai Sự sống

Sau 13,8 tỷ năm cuộc đời, Vũ trụ của chúng ta thức tỉnh và bắt đầu có nhận thức về chính nó. Từ một hành tinh nhỏ màu xanh, những phần tử tí hon và có ý thức của Vũ trụ bắt đầu chăm chú quan sát hệ vũ trụ ngoài kia bằng kính viễn vọng và liên tục khám phá ra rằng mọi thứ họ từng nghĩ là tồn tại trên đời lại chỉ là một phần nhỏ của điều còn vĩ đại hơn: hệ mặt trời, thiên hà và vũ trụ với hơn trăm tỷ thiên hà khác được sắp xếp thành một mô hình phức tạp gồm các nhóm, quần và siêu quần thiên hà. Mặc dù những kẻ ngấm sao có năng lực tự nhận thức này vẫn bất đồng quan điểm ở nhiều mặt, nhưng họ đều đồng ý rằng những thiên hà này thực xinh đẹp và đáng kinh ngạc.

Nhưng cái đẹp nằm trong mắt người thường thức, không phải trong các định luật vật lý; vì vậy, trước khi Vũ trụ của chúng ta thức tỉnh, vẻ đẹp không tồn tại. Điều này khiến cho sự kiện thức tỉnh của vũ trụ chúng ta càng trở nên tuyệt vời và đáng để ngợi ca hơn nữa: Vũ trụ từ một con thây ma vô thức lại biến đổi thành một hệ sinh thái sống nuôi dưỡng tinh thần tự nhìn lại bản thân, nuôi dưỡng vẻ đẹp và hy vọng – cùng nỗ lực theo đuổi mục tiêu, ý nghĩa và lý tưởng. Nếu Vũ trụ của chúng ta không bao giờ thức dậy thì, theo tôi, nó hoàn toàn vô dụng – nó sẽ chỉ là một khoảng không gian lãng phí khổng lồ. Nếu Vũ trụ của chúng ta lại trở về với giấc ngủ vĩnh viễn do một số thảm họa vũ trụ hoặc rủi ro tự gây ra, thì than ôi, nó lại trở nên vô nghĩa.

Mặt khác, mọi thứ sẽ có thể hay ho hơn nữa. Chúng ta vẫn chưa biết liệu loài người có phải là những kẻ ngấm sao đầu tiên hay duy nhất trong vũ trụ không, nhưng chúng ta đã nghiên cứu đủ về Vũ trụ

để biết rằng cho tới giờ có thể nó vẫn chưa thức tỉnh hoàn toàn. Có lẽ chúng ta giống như ánh nhận thức đầu tiên lóe lên, khi ta vừa tỉnh dậy từ giấc ngủ sớm mai: đó là dự cảm về một nhận thức rõ ràng hơn sẽ đến khi ta mở mắt và hoàn toàn tỉnh giấc. Có lẽ sự sống sẽ lan rộng khắp vũ trụ của chúng ta và sinh sôi hàng tỷ hoặc tỷ tỷ năm nữa – và có lẽ điều đó sẽ phụ thuộc vào các quyết định mà chúng ta đưa ra ở đây, trên hành tinh nhỏ bé này và ở thời đại này.

Lược sử về Sự phức tạp

Vậy sự thức tỉnh tuyệt vời này đã xảy ra thế nào? Đó không phải là một sự kiện độc lập mà chỉ là một bước trong cả quá trình 13,8 tỷ năm không ngừng nghỉ, điều khiến cho Vũ trụ của chúng ta trở nên càng phức tạp và thú vị hơn – với tốc độ tăng dần nhanh hơn bao giờ hết.

Là một nhà vật lý, tôi cảm thấy may mắn khi có thể dành phần lớn thời gian trong một phần tư thế kỷ qua để giúp xác định lịch sử vũ trụ; đó quả là một hành trình khám phá tuyệt vời. Từ ngày tôi còn là nghiên cứu sinh, chúng tôi đã đi từ việc tranh luận liệu Vũ trụ của chúng ta 10 hay 20 tỷ năm tuổi tới việc tranh luận xem nó đã 13,7 hay 13,8 tỷ năm tuổi, nhờ có tổ hợp của kính viễn vọng tốt hơn, máy tính xịn hơn và hiểu biết sâu hơn. Các nhà vật lý như tôi vẫn chưa biết chắc điều gì đã gây ra Vụ Nổ Lớn (Big Bang), rằng liệu đây có thực là khởi đầu của mọi thứ hay chỉ là phần tiếp theo của một giai đoạn trước đó. Tuy nhiên, chúng tôi đã thu được hiểu biết khá chi tiết về những gì diễn ra kể từ Vụ Nổ Lớn, nhờ một loạt các phép đo chất lượng cao, vì vậy hãy cho tôi vài phút để tóm tắt 13,8 tỷ năm lịch sử vũ trụ.

Ban đầu ánh sáng tồn tại. Vào khoảnh khắc đầu tiên sau Vụ Nổ Lớn, toàn bộ phần không gian mà kính thiên văn của chúng ta có thể quan sát được trên lý thuyết (“Vũ trụ quan sát được của chúng ta”, hay “Vũ trụ của chúng ta” cho ngắn gọn) nóng hơn và sáng hơn lõi Mặt trời rất nhiều và nó giãn nở nhanh chóng. Mặc dù điều này nghe có vẻ “hoàn hảo”, nhưng nó cũng tẻ nhạt khi xét đến khía cạnh rằng Vũ trụ của chúng ta cũng chỉ là một món súp các hạt cơ bản không có sự sống, dày đặc, nóng và trông đồng dạng đến nhàm

chán. Mọi thứ dường như đều khá giống nhau và cấu trúc thú vị duy nhất tạo bởi các sóng âm thanh yếu ớt mang hình dáng ngẫu nhiên khiến cho món súp đặc hơn 0,001% ở đôi chỗ. Nhiều người cho là những sóng yếu ớt này sinh ra từ thứ mang tên thăng giáng lượng tử, vì nguyên lý bất định của Heisenberg trong cơ học lượng tử không chấp nhận bất cứ thứ gì hoàn toàn nhàm chán và đồng dạng.

Khi Vũ trụ của chúng ta giãn nở và hạ nhiệt, nó trở nên thú vị hơn khi các hạt kết hợp thành các vật thể ngày càng phức tạp. Trong khoảnh khắc đầu tiên, lực hạt nhân mạnh đã nhóm các hạt quark thành proton (hạt nhân hiđro) và neutron, một số trong đó lần lượt hợp nhất thành hạt nhân heli trong vòng vài phút. Khoảng 400.000 năm sau, lực điện từ nhóm các hạt nhân này với các electron để tạo ra những nguyên tử đầu tiên. Trong khi Vũ trụ của chúng ta tiếp tục giãn nở, các nguyên tử này dần dần nguội đi, trở thành thứ khí tối lạnh, và đêm tối đầu tiên này của lịch sử đã kéo dài khoảng 100 triệu năm. Đêm dài này đã dẫn đường cho bình minh vũ trụ của chúng ta khi lực hấp dẫn khuếch đại thành công những biến động trong chất khí tối lạnh kia, kéo các nguyên tử lại với nhau để tạo thành các ngôi sao và thiên hà đầu tiên. Những ngôi sao đầu tiên này tạo ra nhiệt và ánh sáng bằng cách tổng hợp hiđro thành các nguyên tử nặng hơn như cacbon, ôxi và silicon. Khi những ngôi sao này chết, các nguyên tử mà chúng tạo ra được vũ trụ tái chế và tạo thành các hành tinh xung quanh các ngôi sao thế hệ thứ hai.

Tới một thời điểm, một nhóm nguyên tử được sắp xếp thành một mô hình phức tạp có thể vừa duy trì vừa tái tạo chính nó. Vì vậy, chẳng bao lâu có hai bản sao và con số cứ tiếp tục nhân đôi. Chỉ mất 40 lần nhân đôi để tạo ra một nghìn tỷ bản sao, do đó sự vật tự sao chép đầu tiên này nhanh chóng trở thành một sức mạnh đáng cân nhắc. Sự sống đã đến.

Ba giai đoạn của sự sống

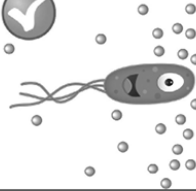
Việc định nghĩa sự sống xưa nay vẫn gây tranh cãi. Có rất nhiều định nghĩa trái ngược, một số trong đó có những yêu cầu rất cụ thể như là phải hình thành từ các tế bào, tức là loại trừ cả máy móc thông minh trong tương lai lẫn các nền văn minh ngoài Trái đất. Bởi

chúng tôi không muốn suy nghĩ của mình về tương lai của sự sống chỉ giới hạn trong các loài mà chúng ta đã biết, nên ta hãy định nghĩa sự sống rộng hơn: như là một quá trình có thể giữ được sự phức tạp và có thể tái tạo. Thứ được tái tạo không phải là vật chất (cấu tạo từ nguyên tử) mà là thông tin (cấu tạo từ các bit) có thể chỉ rõ trình tự sắp xếp các nguyên tử. Khi một vi khuẩn tạo ra bản sao DNA của nó thì nó không tạo ra nguyên tử mới nào, nhưng sẽ có một tập hợp các nguyên tử mới được sắp xếp với thứ tự giống hệt nguyên gốc, nhờ vậy thông tin được sao chép. Nói cách khác, chúng ta có thể xem sự sống như một hệ thống xử lý thông tin tự tái tạo có thông tin (phần mềm) xác định cả hành vi lẫn thiết kế phần cứng của nó.

Giống như bản thân Vũ trụ của chúng ta, sự sống dần trở nên phức tạp và thú vị hơn¹⁰, và tôi phân loại các dạng sự sống thành ba mức độ phức tạp cho thuận tiện: Sự sống 1.0, 2.0 và 3.0. Tôi đã tóm tắt ba cấp độ này trong hình 1.1.

Sự sống lần đầu tiên xuất hiện trong Vũ trụ của chúng ta như thế nào, khi nào và ở đâu vẫn còn là một câu hỏi để ngỏ, nhưng chúng ta có những bằng chứng rõ ràng rằng sự sống lần đầu xuất hiện trên Trái đất khoảng 4 tỷ năm trước. Không lâu sau, hành tinh của chúng ta trở thành một bộ sưu tập đa chủng loại các dạng sống đông đúc. Dạng thành công nhất có khả năng phản ứng trước môi trường theo một vài cách đã nhanh chóng vượt qua số còn lại. Cụ thể, chúng giống những thứ mà các nhà khoa học máy tính gọi là “tác tử thông minh”: chúng là các thực thể thu thập thông tin về môi trường sống từ các cảm biến, rồi xử lý thông tin này để quyết định cách phản ứng với môi trường. Điều này có thể bao gồm việc xử lý thông tin rất phức tạp, như khi ta sử dụng thông tin từ mắt và tai để quyết định sẽ nói gì trong một cuộc đối thoại. Nhưng nó cũng có thể sử dụng phần cứng và phần mềm khá đơn giản.

Ví dụ, nhiều loại vi khuẩn có cảm biến đo nồng độ đường trong chất lỏng xung quanh chúng và có thể bơi bằng các tiên mao có cấu trúc giống động cơ chân vịt. Phần cứng liên kết bộ cảm biến với tiên mao có thể thực hiện thuật toán đơn giản nhưng hữu ích sau đây: “Nếu cảm biến nồng độ đường báo về giá trị thấp hơn vài giây trước, thì nghịch chuyển vòng xoay của tiên mao để tôi thay đổi hướng.”

Nó có thể tồn tại và tái tạo không?			  Hẹn gặp lại!
Nó có thể thiết kế phần mềm không?		  ¡Hola!	  ¡Hola!
Nó có thể thiết kế phần cứng không?	 	  Hi!	  Hi!
	Sự sống 1.0 (sinh học đơn giản)	Sự sống 2.0 (văn hóa)	Sự sống 3.0 (công nghệ)

Hình 1.1: Ba giai đoạn của sự sống: tiến hóa sinh học, tiến hóa văn hóa và tiến hóa công nghệ. Sự sống 1.0 không thể thiết kế lại phần cứng hoặc phần mềm của nó trong suốt vòng đời: cả hai thứ đều được xác định bởi DNA và chỉ thay đổi do tiến hóa qua nhiều thế hệ. Ngược lại, sự sống 2.0 có thể thiết kế lại phần lớn phần mềm của nó: con người có thể học các kỹ năng phức tạp mới – như ngôn ngữ, thể thao hay chuyên môn – và về cơ bản có thể mở rộng thế giới quan và mục tiêu. Sự sống 3.0 vẫn chưa tồn tại trên Trái đất, có khả năng thiết kế lại không chỉ phần mềm mà cả phần cứng của nó, thay vì phải đợi tiến hóa dần dần qua nhiều thế hệ.

Bạn đã học kỹ năng nói cùng vô số kỹ năng khác. Nhưng vi khuẩn thì không giỏi học tập đến thế. DNA của chúng định rõ không chỉ thiết kế phần cứng, chẳng hạn như cảm biến đường và tiên mao, mà còn cả thiết kế phần mềm. Chúng không bao giờ học cách bơi về phía đường; thay vào đó, thuật toán này đã được mã hóa cứng vào DNA của chúng ngay từ đầu. Dĩ nhiên điều này cũng yêu cầu một quá trình học tập, nhưng nó không diễn ra trong vòng đời của riêng con vi khuẩn cụ thể này mà xảy ra trong quá trình tiến hóa trước đó của loài vi khuẩn đó, thông qua quá trình thử-và-sai chậm chạp kéo dài qua nhiều thế hệ; trong quá trình này, lựa chọn tự nhiên ưu tiên những

đột biến DNA ngẫu nhiên có khả năng cải thiện lượng đường tiêu thụ. Một số đột biến này giúp cải thiện thiết kế của tiên mao và các phần cứng khác, trong khi các đột biến còn lại cải thiện hệ thống xử lý thông tin vì khuẩn có chức năng thực hiện thuật toán tìm đường và các phần mềm khác.

Vi khuẩn như vậy là điển hình của cái tôi gọi là “Sự sống 1.0”: sự sống nơi cả phần cứng và phần mềm tiến hóa thay vì được thiết kế. Còn chúng ta là những ví dụ của “Sự sống 2.0”: sự sống với phần cứng tiến hóa, nhưng phần lớn phần mềm được thiết kế. Khi nói đến phần mềm, tôi muốn ngụ ý tất cả thuật toán và kiến thức bạn sử dụng để xử lý thông tin từ các giác quan và quyết định phản hồi – mọi thứ, từ khả năng nhận ra bạn bè khi thấy họ tới khả năng đi đứng, đọc, viết, tính toán, hát và kể chuyện cười.

Bạn không thể thực hiện bất kỳ nhiệm vụ nào trên đây ngay khi bạn sinh ra, vì vậy chúng ta sẽ lập trình các phần mềm này vào bộ não của mình thông qua quá trình học tập. Gia đình và giáo viên là những người thiết kế giáo trình cho bạn thời thơ ấu, nhưng rồi bạn sẽ dần có thêm quyền thiết kế phần mềm của chính mình. Có thể trường học cho phép bạn học tiếng nước ngoài tự chọn: bạn muốn cài đặt vào bộ não một mô-đun phần mềm để bạn nói được tiếng Pháp hay tiếng Tây Ban Nha? Bạn muốn học chơi quần vợt hay cờ vua? Bạn muốn học để trở thành đầu bếp, luật sư hay dược sĩ? Bạn muốn tìm hiểu thêm về trí tuệ nhân tạo (AI) và tương lai của sự sống qua việc đọc một cuốn sách về nó không?

Khả năng thiết kế phần mềm này cho phép Sự sống 2.0 trở nên thông minh hơn nhiều so với Sự sống 1.0. Một trí thông minh cao cấp đòi hỏi lượng lớn cả phần cứng (cấu tạo nguyên tử) và phần mềm (cấu tạo bit). Việc hầu hết phần cứng ở con người chúng ta được thêm vào sau khi sinh (thông qua tăng trưởng) là điều rất hữu ích, bởi nhờ vậy mà kích thước cuối cùng của chúng ta không bị giới hạn theo chiều rộng đường dẫn sinh của mẹ mình. Tương tự, việc phần lớn phần mềm của chúng ta được thêm vào sau khi sinh (thông qua học tập) rất hữu ích, bởi nhờ vậy mà trí tuệ tối thượng của chúng ta không bị giới hạn theo lượng thông tin truyền được tới mình tại thời điểm thụ thai thông qua DNA, theo phong cách 1.0. Tôi nặng gấp khoảng 25 lần so với khi tôi được sinh ra và khớp nối thần kinh liên

kết các nơ-ron tế bào trong não tôi lưu trữ thông tin nhiều hơn 100.000 lần so với DNA mà tôi được sinh ra. Khớp thần kinh lưu trữ tất cả kiến thức và kỹ năng của bạn, tương đương gần 100 terabyte¹¹ thông tin, trong khi DNA của bạn chỉ lưu trữ được khoảng một gigabyte, chỉ vừa đủ để chứa một bộ phim tải xuống từ mạng. Do đó, chuyện một đứa trẻ mới sinh ra đã nói năng thành thạo và có thể làm tốt bài thi vào đại học là bất khả thi về mặt vật lý: thông tin không thể được tải sẵn vào bộ não của bé, vì mô-đun thông tin chính mà bé nhận được từ cha mẹ (DNA của bé) không có đủ dung lượng lưu trữ thông tin.

Khả năng thiết kế phần mềm khiến Sự sống 2.0 không chỉ thông minh hơn Sự sống 1.0 mà còn linh hoạt hơn. Nếu môi trường thay đổi, phiên bản 1.0 chỉ có thể thích nghi bằng cách tiến hóa chậm chạp qua nhiều thế hệ. Mặt khác, Sự sống 2.0 có thể thích ứng gần như ngay lập tức, thông qua cập nhật phần mềm. Ví dụ, một loại vi khuẩn thường xuyên gặp thuốc kháng sinh có thể tiến hóa thành chủng kháng thuốc qua nhiều thế hệ, nhưng cá thể một vi khuẩn sẽ không hề thay đổi hành vi của nó; ngược lại, một cô gái biết mình bị dị ứng đậu phộng sẽ ngay lập tức thay đổi hành vi và bắt đầu tránh đậu phộng. Sự linh hoạt này mang lại cho Sự sống 2.0 lợi thế còn lớn hơn ở quy mô dân số: mặc dù thông tin trong DNA con người chúng ta không tiến hóa gì đáng kể trong 50 nghìn năm qua, nhưng thông tin được lưu trữ tập thể trong bộ não, sách vở và máy tính đã bùng nổ. Bằng cách cài đặt một phần mềm cho phép chúng ta giao tiếp thông qua ngôn ngữ nói phức tạp, chúng ta đảm bảo rằng thông tin hữu ích nhất trữ trong bộ não một người sẽ được sao chép sang bộ não khác, nhờ đó thông tin này có khả năng tồn tại ngay cả khi bộ não nguyên bản chết đi. Bằng cách cài đặt mô-đun phần mềm cho phép chúng ta đọc và viết, chúng ta có thể lưu trữ và chia sẻ nhiều thông tin hơn khả năng ghi nhớ. Bằng cách phát triển những phần mềm não bộ có khả năng sản xuất công nghệ (qua nghiên cứu khoa học và kỹ thuật), chúng ta cho phép nhiều người trên thế giới truy cập phần lớn thông tin trên thế giới chỉ với vài cú nhấp chuột.

Tính linh hoạt này giúp Sự sống 2.0 chiếm ưu thế trên Trái đất. Thoát khỏi xiềng xích di truyền, kiến thức tổng hợp của con người

tiếp tục phát triển với tốc ngày càng tăng khi mỗi bước đột phá lại mở đường cho bước tiếp theo: ngôn ngữ, văn bản, in ấn, khoa học hiện đại, máy tính, mạng Internet, v.v. Bước tiến hóa văn hóa ngày một nhanh nhờ việc chia sẻ phần mềm đã trở thành động lực thống trị, định hình tương lai con người chúng ta, làm cho quy trình tiến hóa sinh học chậm chạp khủng khiếp của chúng ta gần như không còn quan trọng.

Tuy nhiên, bất chấp những công nghệ tiên tiến nhất ngày nay, tất cả các dạng sống mà chúng ta biết vẫn bị giới hạn bởi phần cứng sinh học của chúng. Không ai có thể sống một triệu năm, ghi nhớ tất cả thông tin của Wikipedia, hiểu biết về tất cả các ngành khoa học hoặc bay được vào vũ trụ mà không cần tới tàu vũ trụ. Không ai có thể biến vũ trụ phần lớn là vô hồn của chúng ta thành một không gian sinh quyển đa dạng phát triển tới hàng tỷ hoặc tỷ tỷ năm, giúp cho vũ trụ của chúng ta cuối cùng cũng phát huy hết tiềm năng và hoàn toàn thức tỉnh. Tất cả những điều này đòi hỏi sự sống phải trải qua lần nâng cấp cuối cùng, tiến tới Sự sống 3.0, có khả năng thiết kế không chỉ phần mềm mà cả phần cứng của nó. Nói cách khác, Sự sống 3.0 làm chủ vận mệnh của chính nó, hoàn toàn thoát khỏi xiềng xích tiến hóa.

Ranh giới giữa ba giai đoạn này của sự sống là hơi mờ nhạt. Nếu vi khuẩn là Sự sống 1.0 và con người là Sự sống 2.0, thì bạn có thể phân loại chuột là 1.1: chúng có thể học nhiều thứ, nhưng không đủ để phát triển ngôn ngữ hoặc phát minh ra Internet. Hơn nữa, vì không có ngôn ngữ nên phần lớn những gì chúng học được bị mất đi khi chết chứ không truyền lại cho thế hệ tiếp theo. Tương tự, bạn có thể tranh luận rằng con người ngày nay nên được coi là Sự sống 2.1: chúng ta có thể thực hiện những nâng cấp nhỏ cho phần cứng như trồng răng giả, đầu gối nhân tạo và máy điều hòa nhịp tim, nhưng chưa có những đột phá như việc cao hơn gấp mười lần hoặc có bộ não lớn hơn hàng ngàn lần.

Tóm lại, chúng ta có thể chia sự phát triển của sự sống thành ba giai đoạn, phân biệt bằng khả năng tự thiết kế của sự sống:

- Sự sống 1.0 (giai đoạn sinh học): tiến hóa phần cứng và phần mềm

- Sự sống 2.0 (giai đoạn văn hóa): tiến hóa phần cứng, thiết kế phần lớn phần mềm
- Sự sống 3.0 (giai đoạn công nghệ): thiết kế phần cứng và phần mềm

Sau khoảng 13,8 tỷ năm tiến hóa vũ trụ, sự phát triển đã tăng tốc đáng kể tại đây, trên Trái Đất: Sự sống 1.0 xuất hiện gần bốn tỷ năm trước, Sự sống 2.0 (con người chúng ta) xuất hiện gần 100 thiên niên kỷ trước và rất nhiều nhà nghiên cứu AI dự đoán Sự sống 3.0 có thể sẽ xuất hiện trong thế kỷ tới, có lẽ ngay từ thế hệ chúng ta, được những tiến bộ trong AI kích hoạt. Điều gì sẽ xảy ra và điều đó có ý nghĩa gì với chúng ta? Đó là chủ đề của cuốn sách này.

Tranh cãi

Vấn đề này hết sức gây tranh cãi; các nhà nghiên cứu AI hàng đầu thế giới đều sôi nổi thể hiện bất đồng không chỉ trong các dự báo mà còn trong cả cảm xúc, đi từ lạc quan tự tin đến suy tư nghiêm túc. Họ thậm chí không thể đồng thuận trong những câu hỏi về tác động kinh tế, pháp lý và quân sự trong ngắn hạn của AI, và những bất đồng của họ tăng lên khi chúng ta mở rộng khung thời gian và đặt câu hỏi về

trí tuệ nhân tạo tổng quát

(AGI¹²) - đặc biệt là với AGI đạt tới trình độ con người và hơn thế, trở thành Sự sống 3.0.

Trí tuệ tổng quát

có thể hoàn thành hầu hết mọi mục tiêu, bao gồm cả học tập, ngược lại với trí tuệ hẹp của một chương trình chơi cờ chẳng hạn.

Thật thú vị khi cuộc tranh luận về Sự sống 3.0 xoay quanh không phải một mà là hai câu hỏi riêng biệt: thời điểm và ý nghĩa. Nó sẽ phát sinh vào thời điểm nào (nếu có bao giờ), và ý nghĩa của nó đối với nhân loại là gì? Theo tôi, có ba trường phái tư tưởng riêng biệt cần được xem xét nghiêm túc, bởi trong mỗi trường phái đều có một

số chuyên gia hàng đầu thế giới. Như minh họa trong hình 1.2, tôi nghĩ họ, theo thứ tự, là phái lý tưởng cách mạng số, phái hoài nghi công nghệ và các thành viên của phong trào AI-có-ích. Hãy để tôi giới thiệu bạn với một số nhà đấu tranh hùng hồn nhất trong từng nhóm.

Phái lý tưởng cách mạng số

Hồi còn nhỏ, tôi vẫn tưởng rằng các nhà tỷ phú hay ra vẻ ta đây và ngạo mạn. Khi tôi gặp Larry Page lần đầu tại Google vào năm 2008, anh ấy đã hoàn toàn phá vỡ định kiến này. Thoải mái trong chiếc quần bò và áo sơ-mi trông rất tầm thường, anh hẳn sẽ không khác gì so với những người tham gia một chuyến dã ngoại tại MIT¹³. Phong cách nói nhẹ nhàng chín chắn và nụ cười thân thiện của Larry khiến tôi cảm thấy thư thái chứ không e ngại khi trò chuyện với anh. Ngày 18 tháng 7 năm 2015, chúng tôi lại tình cờ gặp nhau tại một bữa tiệc ở Thung lũng Napa do Elon Musk và người vợ lúc ấy là cô Talulah tổ chức, và nói chuyện với nhau về nỗi ám ảnh những thứ bản thủ của con trẻ. Tôi giới thiệu cho anh cuốn văn học kinh điển sâu sắc The Day My Butt Went Psycho (Ngày cặp mông tôi nổi khùng) của Andy Griffiths và Larry đặt mua nó ngay lập tức. Lúc ấy tôi đã vất vả tự nhắc nhở rằng lịch sử có thể viết về anh như một người có tầm ảnh hưởng nhất mọi thời đại: tôi đoán rằng, nếu sự sống số siêu thông minh chinh phục Vũ trụ của chúng ta trong thời của tôi, thì đó hẳn là do những quyết định của Larry.

Khi nào AI sẽ vượt qua trình độ con người?	Không bao giờ	PHÁI HOÀI NGHI CÔNG NGHỆ				
	Trong 300 năm					
	Trong 100 năm					
	Trong 50 năm	PHÁI BẢN LŨI	PHONG TRÀO AI-CÓ-ÍCH			PHÁI LÝ TƯỞNG CÁCH MẠNG SỐ
	Trong vài thập kỷ					
	Trong vài năm tới	HẦU NHƯ KHÔNG CÓ NGƯỜI NÀO				
	Chắc chắn tệ	Có lẽ tệ	Rất không chắc chắn	Có lẽ tốt	Chắc chắn tốt	
	Nếu AI siêu phẩm xuất hiện thì có phải điều tốt hay không?					

Hình 1.2: Hầu hết các tranh cãi xung quanh trí tuệ nhân tạo mạnh (có thể ngang tài con người ở bất kỳ nhiệm vụ nhận thức nào) đều

xoay quanh hai câu hỏi: Khi nào (nếu có) nó sẽ xảy ra, và liệu việc đó có phải là một điều tốt cho nhân loại? Những người hoài nghi công nghệ và lý tưởng cách mạng số đồng ý rằng chúng ta không nên lo lắng, nhưng bởi những lý do rất khác nhau: nhóm đầu tin rằng trí tuệ nhân tạo tổng quát (AGI) trình độ con người sẽ không xảy ra trong tương lai gần, trong khi nhóm sau nghĩ nó sẽ xảy ra nhưng gần như chắc chắn đó sẽ là việc tốt. Người thuộc phong trào AI-có-ích cảm thấy rằng việc lo ngại là cần thiết và hữu ích, bởi nghiên cứu và thảo luận về an toàn AI hiện nay sẽ làm tăng cơ hội có một kết quả tốt. Những người bàn lùi tin vào hậu quả xấu và phản đối AI. Bảng trên phần nào được truyền cảm hứng từ Tim Urban.¹

Rồi vợ chồng tôi đã dùng bữa tối với Larry và vợ anh, cô Lucy. Chúng tôi thảo luận về việc liệu máy móc có nhất thiết phải có ý thức hay không. Anh ấy tranh luận rằng vấn đề này là một sự đánh lạc hướng. Tối muộn hôm đó, sau khi thưởng thức cocktail, anh và Elon đã tranh luận rất lâu và khì thế về tương lai của AI và những gì nên làm. Tới quá nửa đêm, vòng tròn những người theo dõi cuộc tranh luận tiếp tục mở rộng. Larry bảo vệ nhiệt tình quan điểm mà tôi cho là

chủ nghĩa lý tưởng hóa cách mạng số

: sự sống số là bước tiếp theo hết sức tự nhiên và được kỳ vọng của tiến hóa vũ trụ; nếu chúng ta để trí tuệ kỹ thuật số được tự do thay vì cố gắng ngăn chặn hoặc nô dịch chúng, chúng ta gần như chắc chắn đạt được kết quả tốt. Tôi coi Larry là người diễn giải có ảnh hưởng nhất của chủ nghĩa lý tưởng hóa cách mạng số. Anh lập luận rất lạc quan rằng, nếu sự sống của chúng ta có thể lan rộng khắp thiên hà này hay xa hơn nữa thì sự lan rộng đó sẽ diễn ra dưới dạng kỹ thuật số. Mối quan tâm chính của anh là sự hoang tưởng về AI sẽ trì hoãn thời kỳ lý tưởng cách mạng số và/hoặc dẫn tới việc khiến AI bị quân sự tiếp quản, một tình huống đối nghịch với câu khẩu hiệu “Không làm điều ác” của Google. Elon liên tục phản pháo, yêu cầu Larry làm rõ chi tiết lập luận của mình, chẳng hạn như lý do tại sao anh tự tin rằng sự sống số sẽ không phá hủy mọi thứ mà chúng ta quan tâm. Đôi lần, Larry cáo buộc Elon phân biệt giống loài, rằng Elon đối xử với các dạng sống nhất định như thể chúng thấp kém hơn chỉ vì

chúng làm từ silicon chứ không phải cacbon. Chúng ta sẽ quay trở lại khám phá chi tiết những vấn đề và tranh luận thú vị này , bắt đầu từ chương 4.

Mặc dù Larry dường như bị áp đảo vào cái đêm hè ẩm áp bên hồ bơi ấy, chủ nghĩa lý tưởng hóa cách mạng số mà anh biện giải hùng hồn có rất nhiều người ủng hộ nổi bật. Nhà rô-bốt học theo chủ nghĩa vị lai Hans Moravec đã truyền cảm hứng cho cả một thế hệ lý tưởng cách mạng số với tác phẩm kinh điển *Mind Children* (Những đứa con trí óc) năm 1988. Truyền thống này lại được nhà phát minh Ray Kurzweil tiếp tục và trau chuốt. Richard Sutton, một trong những người tiên phong ở nhánh con của lĩnh vực AI mang tên học tăng cường, đã bảo vệ nhiệt tình chủ nghĩa lý tưởng hóa cách mạng số tại hội nghị Puerto Rico mà tôi sẽ sớm tường thuật cho bạn.

Phái hoài nghi công nghệ

Một nhóm các nhà tư tưởng nổi bật khác cũng không lo lắng về AI, nhưng vì một lý do hoàn toàn khác: họ nghĩ rằng việc xây dựng AGI siêu phàm khó đến nỗi nó sẽ không xảy ra trong hàng trăm năm tới, do đó họ xem việc lo lắng về nó từ bây giờ là ngớ ngẩn. Tôi gọi quan điểm này là hoài nghi công nghệ, được Andrew Ng trình bày hùng hồn: “Lo sợ về sự xuất hiện của người máy sát thủ cũng giống như lo lắng về bùng nổ dân số trên sao Hỏa.” Andrew là nhà khoa học chính tại Baidu, Google của Trung Quốc, gần đây ông đã lặp lại lập luận này khi tôi nói chuyện cùng ông tại một hội nghị ở Boston. Ông cũng bày tỏ mối lo ngại trước nguy cơ AI sẽ là sự phân tâm có khả năng gây hại, làm chậm tiến trình AI. Thái độ tương tự cũng được thể hiện bởi những người hoài nghi công nghệ khác như Rodney Brooks, cựu giáo sư MIT phát minh ra rô-bốt hút bụi Roomba và người máy công nghiệp Baxter. Mặc dù những người lý tưởng cách mạng số và những người hoài nghi công nghệ đồng ý rằng chúng ta không nên lo lắng về AI, nhưng thật thú vị là họ hiếm có điểm chung nào khác. Hầu hết những người lý tưởng cách mạng số nghĩ rằng AGI có trình độ con người có thể được phát minh trong vòng 20 đến 100 năm tới. Những người hoài nghi công nghệ coi đây là giấc mộng viễn vông thiếu hiểu biết và thường chế nhạo điểm kỳ dị

công nghệ¹⁴ trong tiên đoán trên là “sự u mê của lũ đầu to mắt cận”. Khi tôi gặp Rodney Brooks tại một bữa tiệc sinh nhật vào tháng 12 năm 2014, anh ấy nói với tôi rằng chắc chắn 100% điều này sẽ không xảy ra trong thời của tôi. Tôi hỏi trong một email tiếp theo: “Anh có chắc ý anh không phải là 99%?” Anh đã trả lời: “Không phải e dè mà nói 99%. 100%. Sẽ không xảy ra đâu.”

Phong trào AI-có-ích

Lần đầu tiên tôi gặp Stuart Russell tại một quán cà phê ở Paris vào tháng 6 năm 2014, anh khiến tôi có ấn tượng của một quý ông người Anh điển hình. Hùng hồn nhưng chu đáo và ăn nói nhẹ nhàng, lại có ánh phiêu lưu trong mắt, với tôi, anh dường như là hóa thân hiện đại của Phileas Fogg, anh hùng thời thơ ấu của tôi từ cuốn tiểu thuyết kinh điển năm 1873 của Jules Verne,

80 ngày vòng quanh thế giới.

Mặc dù anh là một trong những nhà nghiên cứu AI nổi tiếng nhất đương còn sống, từng là đồng tác giả cuốn sách giáo khoa tiêu chuẩn về chủ đề này, sự khiêm tốn và ầm áp của anh nhanh chóng khiến tôi thấy thoải mái. Anh đã giải thích cho tôi rằng tiến độ phát triển của AI đã thuyết phục anh rằng AGI trình độ con người trong thế kỷ này khả thi như thế nào, và mặc dù vẫn hy vọng nhưng anh vẫn không chắc chắn về kết quả tích cực. Có những câu hỏi trọng yếu mà chúng ta cần giải quyết trước, và chúng khó tới mức chúng ta nên bắt đầu nghiên cứu về chúng ngay bây giờ để có sẵn câu trả lời khi cần.

Ngày nay, quan điểm của Stuart khá phổ biến và nhiều nhóm trên thế giới đang theo đuổi trường phái nghiên cứu an toàn AI mà anh ủng hộ. Nhưng trước đó không phải lúc nào cũng vậy. Một bài báo trên tờ Washington Post cho rằng 2015 là năm mà nghiên cứu về an toàn AI trở nên phổ biến. Trước đó, đối thoại về rủi ro AI thường bị hiểu lầm bởi các nhà nghiên cứu AI chính thống và không được coi trọng, như thể đó là sự kích động sợ hãi của phe bàn lùi nhằm cản trở tiến trình AI. Như chúng ta sẽ khám phá trong chương 5, hơn nửa

thế kỷ trước, nhà tiên phong trong ngành máy tính Alan Turing và nhà toán học Irving J. Good, cộng sự của Turing trong quá trình phá giải mật mã Đức thời Thế chiến II, đã đưa ra quan ngại tương tự với Stuart. Trong thập kỷ qua, nghiên cứu về các chủ đề như vậy chủ yếu được thực hiện bởi một số ít các nhà tư tưởng độc lập không nghiên cứu AI chuyên nghiệp, ví dụ như Eliezer Yudkowsky, Michael Vassar và Nick Bostrom. Công trình của họ không mấy ảnh hưởng lên hầu hết các nhà nghiên cứu AI chính thống, những người thường có xu hướng tập trung vào nhiệm vụ hàng ngày là làm cho hệ thống AI thêm thông minh hơn là suy ngẫm về những hậu quả lâu dài sau thành công. Trong số các nhà nghiên cứu AI mà tôi biết đang có mối quan ngại, nhiều người do dự nói ra vì sợ bị coi là những kẻ căm ghét công nghệ thích gieo hoang mang.

Tôi cảm thấy rằng sự chia rẽ này cần được thay đổi, để toàn bộ cộng đồng AI có thể tham gia và tác động đến cuộc đối thoại về cách xây dựng AI có lợi. May thay, tôi không đơn độc. Mùa xuân năm 2014, tôi thành lập một tổ chức phi lợi nhuận mang tên Viện Tương lai Sự sống (Future of Life Institute; FLI; <http://futureoflife.org>) cùng vợ tôi, Meia, ông bạn – nhà vật lý học Anthony Aguirre, sinh viên cao học Harvard Viktoriya Krakovna và người sáng lập Skype, Jaan Tallinn. Mục tiêu của chúng tôi rất đơn giản: giúp đảm bảo rằng tương lai của sự sống sẽ tồn tại và sẽ tuyệt vời nhất có thể. Cụ thể, chúng tôi cảm thấy rằng công nghệ mang lại cho sự sống sức mạnh hoặc để phát triển mạnh mẽ hơn bao giờ hết hoặc tự hủy diệt, và chúng tôi thích vẽ đầu hơn.

Cuộc họp đầu tiên của chúng tôi là một buổi hội ý tại nhà vào ngày 15 tháng 3 năm 2014, với khoảng 30 sinh viên, giáo sư và các nhà tư tưởng khác ở khu vực Boston. Chúng tôi đã đồng thuận rộng rãi rằng mặc dù chúng ta nên chú ý đến công nghệ sinh học, vũ khí hạt nhân và biến đổi khí hậu nhưng mục tiêu chính đầu tiên của chúng ta nên là giúp phổ biến nghiên cứu về an toàn AI. Đồng nghiệp môn vật lý tại MIT của tôi, Frank Wilczek, người đã giành giải thưởng Nobel nhờ phát hiện ra hành vi của hạt quark, đề xuất khởi động bằng cách viết một bài luận trên báo để thu hút chú ý và khiến người ta khó bỏ qua vấn đề này hơn. Tôi đã liên hệ với Stuart Russell (lúc đó tôi vẫn chưa gặp) và đồng nghiệp môn vật lý Stephen Hawking;

họ đều đồng ý tham gia cùng tôi và Frank với tư cách đồng tác giả. Sau nhiều lần chỉnh sửa, bài luận của chúng tôi bị tờ New York Times và nhiều tờ báo Mỹ khác từ chối, vì vậy chúng tôi đã đăng nó lên tài khoản blog Huffington Post của tôi. Hạnh phúc thay, Arianna Huffington đã gửi email phản hồi: “vui mừng vì có nó! Chúng tôi sẽ đăng ở trang #1!”, và vị trí trên cùng trang nhất này đã kích động một làn sóng truyền thông về an toàn AI kéo dài đến hết năm, cùng sự tham gia của Elon Musk, Bill Gates và các nhà lãnh đạo công nghệ khác. Cuốn sách Superintelligence (Siêu trí tuệ) của Nick Bostrom đã phát hành vào mùa thu năm đó và thúc đẩy mạnh hơn cho cuộc tranh luận đang lớn dần trong cộng đồng.

Mục tiêu tiếp theo cho chiến dịch AI-có-ích của FLI là mời các nhà nghiên cứu AI hàng đầu thế giới đến một hội nghị để xóa bỏ hiểu lầm, xác lập sự đồng thuận và xây dựng các kế hoạch hữu dụng. Chúng tôi biết rằng sẽ rất khó để thuyết phục một đám đông lừng lầy như vậy đến một hội nghị do những kẻ ngoại đạo lạ hoắc tổ chức, đặc biệt là với chủ đề gây tranh cãi như thế, vì vậy chúng tôi đã cố gắng hết sức có thể: chúng tôi cấm phương tiện truyền thông tham gia, chọn địa điểm ở khu nghỉ mát bãi biển vào tháng Một (ở Puerto Rico), tổ chức miễn phí (nhờ Jaan Tallinn đã hào phóng) và đặt cho hội nghị một cái tên ít gây lo ngại nhất mà chúng tôi có thể nghĩ được: “Tương lai của AI: Cơ hội và Thách thức”. Quan trọng nhất, chúng tôi đã có thể hợp tác với Stuart Russell, nhờ anh nên chúng tôi đã mở rộng được ban tổ chức, có sự tham gia của các nhà lãnh đạo ngành AI từ cả mảng học thuật và công nghiệp – có Demis Hassabis từ DeepMind của Google, người sau này cho thấy AI có thể đánh bại con người trong môn cờ vây. Càng biết về Demis, tôi càng nhận ra anh có tham vọng không chỉ làm cho AI mạnh mẽ, mà còn làm cho AI có ích.

Kết quả là một cuộc họp đáng chú ý của những trí tuệ (hình 1.3). Các nhà nghiên cứu AI đã tham gia cùng các nhà kinh tế học, học giả pháp lý, các nhà lãnh đạo công nghệ hàng đầu (bao gồm cả Elon Musk) và các nhà tư tưởng khác (bao gồm Vernor Vinge, người đặt ra thuật ngữ “điểm kỳ dị”, trọng tâm của chương 4). Kết quả đã vượt qua cả kỳ vọng lạc quan nhất của chúng tôi. Có lẽ đó là nhờ ánh nắng mặt trời nơi này hòa quyện với rượu vang, hoặc có lẽ chỉ là

nhờ thời điểm thích hợp: mặc dù chủ đề rất gây tranh cãi nhưng tất cả đã đồng thuận đến đáng ngạc nhiên; chúng tôi đã ký kết điều này bằng một bức thư ngỏ² có hơn 8.000 người ký, trong đó có cả một vị cực nổi tiếng trong ngành AI. Ý chính của bức thư là mục tiêu của AI cần được định nghĩa lại: mục tiêu không nên là tạo ra trí tuệ không có định hướng, mà là tạo ra trí tuệ có ích. Bức thư cũng đề cập đến một danh sách chi tiết các chủ đề nghiên cứu mà những người tham gia hội nghị đồng tình là sẽ giúp phát triển mục tiêu này. Phong trào AI-có-ích bắt đầu trở nên phổ biến từ đó. Chúng ta sẽ theo dõi tiến trình của nó ở phần sau cuốn sách.



Hình 1.3: Hội nghị tháng Giêng 2015 tại Puerto Rico tập hợp một nhóm tiêu biểu các nhà nghiên cứu trong ngành AI và các lĩnh vực liên quan. Hàng sau cùng, từ trái sang phải: Tom Mitchell, Seán Ó hÉigeartaigh, Huw Price, Shamil Chandaria, Jaan Tallinn, Stuart Russell, Bill Hibbard, Blaise Agüera y Arcas, Anders Sandberg, Daniel Dewey, Stuart Armstrong, Luke Muehlhauser, Tom Dietterich, Michael Osborne, James Manyika, Ajay Agrawal, Richard Mallah, Nancy Chang, Matthew Putman. Những người đang đứng khác, từ trái sang phải: Marilyn Thompson, Rich Sutton, Alex Wissner-Gross, Sam Teller, Toby Ord, Joscha Bach, Katja Grace, Adrian Weller, Heather Roff-Perkins, Dileep George, Shane Legg, Demis Hassabis, Wendell Wallach, Charina Choi, Ilya Sutskever, Kent Walker, Cecilia Tilli, Nick Bostrom, Erik Brynjolfsson, Steve Crossan, Mustafa Suleyman, Scott Phoenix, Neil Jacobstein, Murray Shanahan, Robin Hanson, Francesca Rossi, Nate Soares, Elon Musk, Andrew McAfee, Bart Selman, Michele Reilly, Aaron VanDevender, Max Tegmark,

Margaret Boden, Joshua Greene, Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel, James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva-Melissa Tez, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. Ngồi: Sam Harris, Tomaso Poggio, Marin Soljai, Viktoriya Krakovna, Meia Chita-Tegmark. Sau máy ảnh: Anthony Aguirre (được chỉnh sửa vào ảnh nhờ trí tuệ trình độ con người ngồi ngay cạnh anh).

Một bài học quan trọng khác từ hội nghị là: các câu hỏi đặt ra từ thành công của AI không chỉ hấp dẫn về trí tuệ mà còn rất quan trọng về mặt đạo đức, bởi các lựa chọn của chúng ta có khả năng ảnh hưởng đến toàn bộ tương lai của sự sống. Các lựa chọn trong quá khứ của nhân loại đôi khi có tầm vóc rất vĩ đại, nhưng luôn bị giới hạn: chúng ta đã hồi phục sau cả những bệnh dịch lớn nhất và những đế quốc vĩ đại nhất rồi cũng sụp đổ. Các thế hệ trước biết rằng, cũng chắc chắn như Mặt Trời phải mọc vào ngày mai, tương lai của con người cũng như vậy, phải giải quyết những tai họa lâu dài như đói nghèo, bệnh tật và chiến tranh. Nhưng một số diễn giả ở Puerto Rico tranh luận rằng lần này sẽ khác: họ cho rằng đây sẽ là lần đầu tiên chúng ta có thể xây dựng công nghệ đủ mạnh để chấm dứt những tai họa này vĩnh viễn – hoặc để kết thúc cả nhân loại. Chúng ta có thể tạo ra những xã hội phát triển mạnh mẽ hơn bao giờ hết, trên Trái đất và có thể xa hơn, hoặc một đế chế giám sát toàn cầu kinh khủng, mạnh đến mức không bao giờ có thể bị lật đổ kiểu như nhà văn Franz Kafka viết ra.



Hình 1.4: Mặc dù các phương tiện truyền thông thường miêu tả Elon Musk là chống đối cộng đồng AI, trên thực tế cả cộng đồng nhất trí rộng rãi rằng nghiên cứu về an toàn AI rất cần thiết. Trong ảnh, vào ngày 4 tháng 1 năm 2015, Tom Dietterich, chủ tịch Hiệp hội vì Sự tiến bộ của Trí tuệ Nhân tạo, cùng Elon bày tỏ phần khích về chương trình nghiên cứu an toàn AI mới mà Elon cam kết tài trợ chỉ vài phút trước đó. Lấp ló sau họ chính là các nhà sáng lập FLI: Meia Chita-Tegmark và Viktoriya Krakovna.

Quan niệm sai lầm

Khi rời Puerto Rico, tôi thực sự tin rằng cuộc đối thoại của chúng tôi về tương lai AI cần được tiếp tục, bởi đó là cuộc trò chuyện quan trọng nhất thời đại của chúng ta.¹⁵ Đó là cuộc trò chuyện về tương lai của tất cả chúng ta, vì vậy nó không nên giới hạn chỉ trong các nhà nghiên cứu AI. Và đây chính là lý do tôi viết cuốn sách này: tôi đã viết nó với hy vọng rằng bạn, độc giả thân mến của tôi, sẽ tham gia cuộc trò chuyện này. Bạn muốn tương lai như thế nào? Chúng ta có nên phát triển vũ khí sát thương tự hành không? Bạn muốn tự động hóa công việc xảy ra theo hướng nào? Bạn sẽ đưa ra lời khuyên nghề nghiệp nào cho trẻ em ngày nay? Bạn thích các công việc mới thay thế cho công việc cũ, hay thích một xã hội không có việc làm, nơi mọi người đều hưởng thụ cuộc sống vui thú và sự giàu có do máy móc tạo ra? Sâu xa hơn, bạn có muốn tạo ra Sự sống 3.0 và phát tán nó ra

vũ trụ của chúng ta không? Chúng ta sẽ kiểm soát máy móc thông minh hay chúng sẽ kiểm soát ta? Liệu các máy móc thông minh sẽ thay thế, cùng tồn tại hay hợp nhất với chúng ta? Làm người trong thời đại trí tuệ nhân tạo thì sẽ thế nào? Bạn muốn nó ra sao, và ta cần làm gì để tạo nên một tương lai như vậy?

Cuốn sách này sẽ giúp bạn tham gia cuộc trò chuyện đó. Như tôi đã đề cập, chúng ta có những tranh cãi thú vị, vì các chuyên gia hàng đầu thế giới chưa đồng thuận với nhau. Nhưng tôi cũng thấy nhiều cuộc tranh cãi giả tạo nhằm chán, trong đó mọi người nói chuyện mà không hiểu nhau. Để giúp chúng ta tập trung vào những tranh cãi thú vị và các câu hỏi mở chứ không phải những hiểu lầm, hãy bắt đầu bằng cách xóa bỏ một số quan niệm nhằm lẫn phổ biến nhất.

Các thuật ngữ như «sự sống», «trí thông minh» và «ý thức» có nhiều định nghĩa khác nhau đều được sử dụng phổ biến, cũng có nhiều quan niệm sai lầm đến từ những người không nhận ra rằng họ đang sử dụng một từ theo hai cách khác nhau. Để đảm bảo rằng bạn và tôi không rơi vào cái bẫy này, tôi đã đặt ra bảng 1.1 để diễn giải cách tôi định nghĩa các thuật ngữ chính trong cuốn sách. Một số định nghĩa sẽ chỉ được giới thiệu và giải thích chính xác trong các chương sau. Xin lưu ý rằng tôi không tuyên bố định nghĩa của tôi đúng hơn các định nghĩa khác – tôi chỉ muốn tránh nhầm lẫn bằng cách làm rõ ý của mình. Bạn sẽ thấy tôi thường định nghĩa rất rộng nhằm tránh thiên vị cho con người, giúp các khái niệm có thể áp dụng cho cả máy móc lẫn con người. Bây giờ hãy đọc bảng thuật ngữ, và quay lại kiểm tra sau này nếu bạn thấy bối rối trước cách tôi sử dụng một trong những từ đó – đặc biệt là trong các chương 4-8.

Bảng thuật ngữ

<i>Sự sống</i>	<i>Quá trình có thể giữ lại sự phức tạp của nó và tái tạo</i>
<i>Sự sống 1.0</i>	<i>Sự sống tiến hóa phần cứng và phần mềm (giai đoạn sinh học)</i>
<i>Sự sống 2.0</i>	<i>Sự sống tiến hóa phần cứng nhưng thiết kế phần lớn phần mềm của nó (giai đoạn văn hóa)</i>
<i>Sự sống 3.0</i>	<i>Sự sống tự thiết kế phần cứng và phần</i>

	mềm (giai đoạn công nghệ)
Trí tuệ	Khả năng hoàn thành các mục tiêu phức tạp
Trí tuệ nhân tạo (AI)	Trí tuệ không ở dạng sinh học
Trí tuệ hẹp	Khả năng hoàn thành một nhóm mục tiêu hẹp, ví dụ chơi cờ hoặc lái xe
Trí tuệ tổng quát	Khả năng hoàn thành gần như mọi mục tiêu, bao gồm cả học tập
Trí tuệ phổ quát	Khả năng thu nạp trí tuệ tổng quát nếu được truy cập vào dữ liệu và tài nguyên
Trí tuệ Nhân tạo Tổng quát [trình độ con nhận thức ở mức ít nhất là ngang bằng con người] (AGI)	Khả năng hoàn thành mọi nhiệm vụ người
AI Trình độ con người	AGI
AI mạnh	AGI
Siêu trí tuệ	Trí tuệ tổng quát vượt xa trình độ con người
Văn minh	Nhóm các dạng sống thông minh tương tác với nhau
Ý thức	Trải nghiệm chủ quan
Cảm chất	Các trải nghiệm chủ quan riêng biệt
Đạo đức	Các nguyên tắc chi phối cách chúng ta cư xử
Mục đích luận	Giải thích về mọi sự vật trong tương quan với mục tiêu hoặc mục đích của chúng thay vì nguyên nhân
Hành vi có mục tiêu	Hành vi dễ dàng giải thích thông qua tác động của nó hơn là nguyên nhân
Có mục tiêu	Biểu lộ hành vi có mục tiêu
Có mục đích	Phục vụ các mục tiêu của bản thân hoặc của một thực thể khác
AI thân thiện	Siêu trí tuệ có mục tiêu phù hợp với mục

	<i>tiêu của chúng ta</i>
<i>Cyborg (Sinh vật cơ khí hóa)</i>	<i>Dạng sống lai giữa con người và máy móc</i>
<i>Bùng nổ trí tuệ</i>	<i>Sự tự cải tiến để quy để nhanh chóng dẫn tới siêu trí tuệ</i>
<i>Điểm kỳ dị</i>	<i>Bùng nổ trí tuệ</i>
<i>Vũ trụ</i>	<i>Vùng không gian mà ánh sáng đã đi qua để chạm tới chúng ta trong suốt 13.8 tỷ năm kể từ Vụ Nổ Lớn</i>

Bảng 1.1: Rất nhiều hiểu lầm về AI nảy sinh khi người ta sử dụng các từ trên với những ý nghĩa khác nhau. Trên đây là định nghĩa tôi áp dụng ở cuốn sách này. (Một số khái niệm sẽ chỉ được giới thiệu và giải thích đầy đủ trong các chương sau.)

Ngoài những nhầm lẫn về thuật ngữ, tôi cũng thấy nhiều cuộc đối thoại về AI bị lạc đề bởi những quan niệm sai lầm đơn giản. Hãy làm sáng tỏ những quan niệm sai lầm phổ biến nhất.

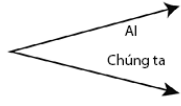
Hiểu lầm về dòng thời gian

Điều đầu tiên liên quan đến dòng thời gian ở hình 1.2: mất bao lâu thì máy móc sẽ thay thế phần lớn AGI trình độ con người? Ở đây, quan niệm sai lầm phổ biến là chúng ta biết khá chắc câu trả lời.

Nhiều người hiểu nhầm rằng chúng ta chắc chắn sẽ có AGI siêu phàm trong thế kỷ này. Trên thực tế, lịch sử công nghệ vốn luôn tràn đầy những cường điệu quá đáng. Các nhà máy điện nhiệt hạch và những chiếc xe bay mà chúng ta từng được hứa sẽ có nay đang nơi đâu? AI cũng bị cường điệu quá nhiều lần trong quá khứ, thậm chí bởi một số người sáng lập của lĩnh vực này: có lần John McCarthy (người đặt ra thuật ngữ “trí tuệ nhân tạo”), Marvin Minsky, Nathaniel Rochester và Claude Shannon đã viết ra một dự báo quá mức lạc quan về những gì có thể đạt được trong hai tháng với máy tính thời kỳ đồ đá như thế này: “Chúng tôi đề xuất tổ chức cuộc nghiên cứu về trí tuệ nhân tạo trong 2 tháng, với 10 người trong mùa hè năm 1956 tại trường Dartmouth... Chúng ta sẽ tìm cách làm cho máy móc sử dụng ngôn ngữ, hình thành năng lực trừu tượng và khái niệm, giải quyết các loại vấn đề hiện đang dành riêng cho con người và tự cải thiện bản thân chúng. Chúng tôi nghĩ sẽ đạt tiến bộ đáng kể trong

một hoặc nhiều vấn đề này nếu để một nhóm các nhà khoa học được lựa chọn cẩn thận làm việc cùng nhau cả mùa hè.”

Mặt khác, còn có quan niệm nhằm lẫn ngược lại nhưng cũng phổ biến là chúng ta chắc chắn sẽ không có AGI siêu phàm trong thế kỷ này. Các nhà nghiên cứu đã đưa ra một loạt ước tính về khoảng cách giữa chúng ta với AGI siêu phàm, nhưng chắc chắn chúng ta không thể nói tự tin và mạnh mẽ rằng xác suất đó sẽ bằng 0 vào thế kỷ này, qua lịch sử kết quả dự đoán thảm hại của phe hoài nghi công nghệ. Một ví dụ là vào năm 1933, Ernest Rutherford, được cho là nhà vật lý hạt nhân vĩ đại nhất thời đại ông, đã nói – chưa đầy 24 giờ trước khi Leo Szilard phát minh ra chuỗi phản ứng hạt nhân – rằng năng lượng hạt nhân là “thứ nhảm nhí” và năm 1956, nhà Thiên văn Hoàng gia Richard Woolley gọi việc du lịch không gian là “toàn rác rưởi”. Hình thức cực đoan nhất của kiểu hiểu lầm này là cho rằng vì bất khả thi về mặt vật lý nên AGI siêu phàm sẽ không bao giờ nằm trong tầm với của chúng ta. Tuy nhiên, các nhà vật lý đều biết rằng một bộ não gồm các hạt quark và electron được sắp xếp để hoạt động như một chiếc máy tính mạnh, và rằng không có định luật vật lý nào ngăn cản chúng ta tạo ra các khối quark còn thông minh hơn nữa.

Hiểu lầm: Siêu trí tuệ vào năm 2100 là không thể tránh khỏi.		Thực tế: Nó có thể xảy ra trong vài thập kỷ tới, vài thế kỷ tới, hoặc không bao giờ: các chuyên gia AI đều bất đồng quan điểm còn chúng ta thì không biết.	
Hiểu lầm: Chỉ những người bàn lùi mới lo lắng về AI.		Thực tế: Rất nhiều nhà nghiên cứu AI hàng đầu cũng lo lắng.	
Lo lắng sai lầm: AI trở nên tàn ác. Lo lắng sai lầm: AI trở nên có ý thức.		Lo lắng thực tế: AI trở nên có năng lực, có mục tiêu lệch lạc với chúng ta.	
Hiểu lầm: Rô-bốt là mối lo ngại chính.		Thực tế: Trí tuệ lệch lạc là mối lo ngại chính: nó không cần bất cứ ai, mà chỉ cần kết nối Internet.	
Hiểu lầm: AI không thể điều khiển con người.		Thực tế: Trí tuệ dẫn tới sự điều khiển: nhờ có trí khôn nên chúng ta mới kiểm soát được loài hổ.	
Hiểu lầm: Máy móc không có mục tiêu.		Thực tế: Tên lửa tâm nhiệt có mục tiêu.	
Lo lắng sai lầm: Siêu trí tuệ chỉ còn cách ta vài năm.	HOẢNG LOẠN	Lo lắng thực tế: Ít nhất còn phải vài thập kỷ nữa, nhưng cũng sẽ cần ngăn dấy thời gian để khiến nó an toàn.	CẦN LÊN KẾ HOẠCH!

Hình 1.5: Hiểu lầm phổ biến về AI siêu trí tuệ.

Đã có không ít khảo sát hỏi ý kiến các nhà nghiên cứu AI rằng sau bao lâu nữa chúng ta sẽ có AGI ở trình độ con người với xác suất ít nhất là 50%; tất cả đều đi đến một kết luận: vì các chuyên gia hàng đầu thế giới vẫn còn bất đồng quan điểm, nên chúng ta không thể biết. Ví dụ, trong cuộc thăm dò ý kiến các nhà nghiên cứu AI tại hội nghị AI Puerto Rico, câu trả lời trung bình là vào năm 2055, nhưng một số nhà nghiên cứu đã đưa ra dự đoán hàng trăm năm nữa hoặc hơn.

Một quan niệm sai lầm khác cũng khá liên quan là những người lo lắng về AI nghĩ rằng nó chỉ còn cách vài năm nữa thôi. Trên thực

tế, hầu hết những người quan ngại về AGI siêu phàm được hỏi đều đoán rằng nó vẫn còn cách ta ít nhất vài thập kỷ nữa. Nhưng họ lập luận rằng, chừng nào chúng ta còn không chắc chắn 100% về khả năng xuất hiện AGI siêu phàm trong thế kỷ này thì tốt nhất ta nên bắt đầu nghiên cứu về tính an toàn ngay bây giờ để chuẩn bị cho tình huống có thể xảy ra. Như chúng ta sẽ thấy trong cuốn sách này, nhiều vấn đề về an toàn khó đến nỗi có thể cần nhiều thập kỷ để giải quyết, nên việc nghiên cứu chúng ngay bây giờ sẽ tốt hơn thay vì đợi đến cái đêm trước khi các tay lập trình viên mãi nốc Bò Húc quyết định kích hoạt AGI trình độ con người.

Hiểu lầm gây tranh cãi

Quan niệm sai lầm phổ biến khác cho rằng những người duy nhất chất chứa lo ngại về AI và ủng hộ nghiên cứu an toàn AI là những người bần lười không biết nhiều về AI. Khi Stuart Russell đề cập đến điều này trong bài phát biểu ở Puerto Rico, tất cả khán giả đều cười ồ. Một quan niệm sai lầm có liên quan ở đây là hỗ trợ nghiên cứu an toàn AI sẽ gây nhiều tranh cãi. Trên thực tế, mọi người không cần tin vào khả năng rủi ro cao thì mới đầu tư một khoản khiêm tốn vào nghiên cứu an toàn AI; thật ra nó cũng không đáng kể, mà chỉ như khoản đầu tư khiêm tốn vào bảo hiểm nhà cửa khi ta tin vào một xác suất cháy nhà không đáng kể.

Theo tôi thấy, chính các phương tiện truyền thông đã làm cho cuộc tranh luận về an toàn AI dường như gây nhiều tranh cãi hơn thực tế. Dù sao thì, tin giật gân thì mới hấp dẫn, và các bài báo trích dẫn không kèm hoàn cảnh để tuyên bố sự diệt vong sắp xảy ra có thể đem lại nhiều cú nhấp chuột hơn so với những câu nhẹ nhàng và cân bằng. Kết quả là, hai người chỉ biết tới quan điểm của nhau nhờ truyền thông trích dẫn có xu hướng nghĩ rằng họ bất đồng với nhau hơn là thực tế. Ví dụ, một người hoài nghi về công nghệ chỉ biết về quan điểm của Bill Gates qua một tờ báo lá cải của Anh có thể nhầm tưởng rằng ông tin siêu trí tuệ sắp xảy ra. Tương tự như vậy, một người trong phong trào AI-có-ích chỉ biết đến trích dẫn từ Andrew Ng về tình trạng quá tải dân số trên Sao Hỏa mà không biết gì khác có thể nhầm tưởng rằng anh không quan tâm đến an toàn AI. Trên thực

tế, cá nhân tôi biết anh có quan tâm – chỉ là ước tính thời gian của anh ấy dài hơn, nên Andrew có xu hướng ưu tiên những thách thức AI ngắn hạn hơn là dài hạn.

Hiểu lầm về rủi ro

Tôi đã đảo mắt chán nản khi nhìn thấy tiêu đề sau trên tờ Daily Mail:³ “Stephen Hawking Cảnh Báo Sự Trỗi Dậy Của Rô-bốt Có Thể Trở Thành Thảm Hoạ Cho Nhân Loại.” Tôi không thể nhớ nổi số lượng bài viết tương tự mà tôi đã thấy. Thông thường, chúng đi kèm với hình ảnh người máy trông ác độc vác vũ khí và cho rằng chúng ta cần lo lắng về việc rô-bốt nổi dậy và giết chết chúng ta bởi chúng sẽ trở nên có ý thức và/hoặc độc ác. Nếu nhìn nhận nhẹ nhàng thì các bài viết như vậy thực sự khá ấn tượng, vì chúng tóm tắt ngắn gọn kịch bản mà các đồng nghiệp làm AI của tôi chẳng lo lắng gì. Kịch bản đó kết hợp đến ba quan niệm sai lầm riêng biệt, lần lượt là: mối quan tâm về ý thức, cái ác và rô-bốt.

Nếu bạn lái xe trên đường, bạn có trải nghiệm chủ quan về màu sắc, âm thanh, v.v.. Nhưng liệu một chiếc xe tự lái có trải nghiệm chủ quan không? Liệu nó có cảm thấy điều gì không khi là chiếc xe tự lái, hay nó giống như một xác sống vô thức không có bất kỳ trải nghiệm chủ quan nào? Mặc dù việc tìm hiểu về bí ẩn này của ý thức cũng rất thú vị và chúng ta sẽ dành cả chương 8 cho nó, nhưng nó không liên quan gì đến nguy cơ AI. Nếu bạn bị một chiếc xe không người lái đâm phải, cảm giác của bạn cũng sẽ không thay đổi dù nó có ý thức chủ quan hay không. Vì thế, thứ ảnh hưởng đến con người chúng ta không phải là cảm xúc chủ quan của AI siêu trí tuệ mà là những gì nó làm.

Nỗi sợ máy móc biến thành ác độc là một chiêu bài đánh lạc hướng khác. Chúng ta không cần lo lắng gì đến ác ý, mà cần lo về năng lực. Theo định nghĩa, AI siêu thông minh sẽ đạt được mục tiêu của nó rất tốt, bất kể mục tiêu đó là gì, vì vậy chúng ta cần đảm bảo rằng mục tiêu của nó phù hợp với chúng ta. Có thể bạn không phải người ghét lũ kiến mà luôn giẫm đạp chúng vì ác ý, nhưng nếu bạn chịu trách nhiệm về một dự án năng lượng xanh thủy điện mà lại có một tổ kiến trong khu vực cần xả nước, thì lũ kiến thật đen đủi.

Phong trào AI-có-ích muốn tránh đưa nhân loại vào vị trí những con kiến đó.

Nhận thức sai lầm về ý thức bước ra từ hiểu lầm rằng máy móc không thể có mục tiêu. Máy móc chắc chắn có mục tiêu, theo nghĩa hẹp là bộc lộ hành vi có mục tiêu: mục tiêu ngắn gọn nhất thể hiện qua hành vi của tên lửa tầm nhiệt là bắn trúng đích. Nếu bạn cảm thấy bị đe dọa bởi một cỗ máy có mục tiêu không phù hợp với bạn thì chính mục tiêu này của nó mới gây phiền não cho bạn, chứ không phải việc liệu cỗ máy có ý thức và có cảm giác về mục đích hay không. Nếu tên lửa tầm nhiệt đó đang đuổi theo bạn, chắc bạn cũng không dám kêu rằng: “Tôi không lo lắng, vì máy móc không thể có mục tiêu!”

Tôi thông cảm với Rodney Brooks và các nhà tiên phong khác trong ngành người máy— những người cảm thấy bị báo chí lá cải bôi xấu bất công; bởi một số nhà báo dường như bị ám ảnh điên cuồng về người máy, họ tô điểm cho nhiều bài báo của mình với các quái vật kim loại đáng hình dử tợn có đôi mắt lóe ánh đỏ. Trên thực tế, mối quan tâm chính của phong trào AI-có-ích không phải là người máy mà chính là trí tuệ: cụ thể là trí tuệ có mục tiêu không phù hợp với chúng ta. Để gây rắc rối cho chúng ta, trí tuệ sai lệch đó không cần cơ thể rô-bốt mà chỉ cần kết nối Internet – trong chương 4, chúng ta sẽ khám phá cách nó thực hiện nhiều vấn đề như vượt mặt thị trường tài chính, phát minh vượt mặt các nhà nghiên cứu là con người, vượt khả năng thao túng các nhà lãnh đạo và phát triển vũ khí mà chúng ta thậm chí chưa thể hiểu được. Ngay cả khi không thể xây dựng người máy về mặt vật lý, một AI siêu-thông-minh và siêu-giàu có thể dễ dàng trả tiền hoặc thao túng không biết bao nhiêu con người để thực hiện mệnh lệnh của nó mà không hay biết, như trong tiểu thuyết khoa học viễn tưởng Neuromancer (Kẻ thao túng trí óc) của William Gibson.

Nhận thức nhầm lẫn về người máy liên quan đến hiểu lầm rằng máy móc không thể kiểm soát con người. Trí tuệ mở đường cho kiểm soát: con người kiểm soát con hổ không phải vì chúng ta mạnh hơn mà bởi chúng ta thông minh hơn. Điều này có nghĩa là nếu chúng ta phải nhường lại vị trí thông minh nhất trên hành tinh của chúng ta, rất có thể chúng ta cũng nhường lại quyền kiểm soát.

Hình 1.5 tóm tắt tất cả những quan niệm sai lầm phổ biến này, để chúng ta có thể vứt bỏ chúng một lần và mãi mãi mà tập trung thảo luận với bạn bè và đồng nghiệp về các vấn đề chính đáng – mà như chúng ta thấy là không hề thiếu!

Con đường phía trước

Trong phần còn lại của cuốn sách, bạn và tôi sẽ cùng nhau khám phá tương lai của sự sống với AI. Hãy thám hiểm chủ đề phong phú và đa diện này theo cách có tổ chức bằng việc khám phá toàn bộ lịch sử sự sống theo khái niệm và theo thứ tự thời gian, sau đó khám phá mục tiêu, ý nghĩa và hành động cần thực hiện để tạo ra tương lai chúng ta muốn.

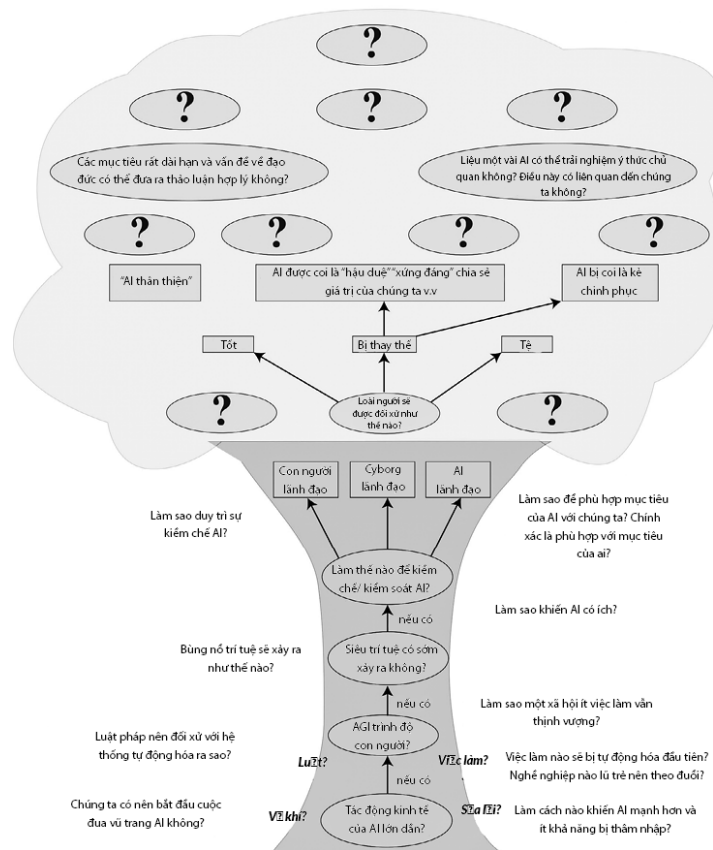
Trong chương 2, chúng ta sẽ khám phá nền tảng của trí tuệ và cách mà vật chất dường như vô tri có thể được sắp xếp lại để ghi nhớ, tính toán và học tập. Khi chúng ta tiến dần về tương lai, câu chuyện của chúng ta sẽ phân nhánh thành nhiều kịch bản, định hướng bằng câu trả lời của một số câu hỏi chính nhất định. Hình 1.6 sẽ tóm tắt các câu hỏi quan trọng mà chúng ta sẽ gặp phải khi lần theo dòng thời gian để đến với thời điểm AI có khả năng ngày càng tiên tiến hơn.

Ngay bây giờ, chúng ta đang phải đối mặt với việc có nên bắt đầu một cuộc chạy đua AI hay không, cùng những vấn đề cần giải đáp để làm cho hệ thống AI ngày sau trở nên ổn định và không bị lỗi. Nếu tác động kinh tế của AI tiếp tục lớn dần, thì chúng ta cũng cần cân nhắc nên hiện đại hóa pháp luật như thế nào và cung cấp những lời khuyên nghề nghiệp gì cho lũ trẻ để chúng có thể tránh các công việc sớm được tự động hóa. Chúng ta sẽ khám phá về những câu hỏi ngắn hạn như vậy trong chương 3.

Nếu AI tiến triển đạt tới trình độ con người thì chúng ta cũng cần chú ý để đảm bảo nó có lợi, và xem xét liệu có thể hay có nên tạo ra một xã hội hưởng thụ hưng thịnh mà không có việc làm hay không. Điều này lại đặt ra thêm câu hỏi rằng liệu AGI sẽ vượt xa trình độ con người bằng sự bùng nổ trí tuệ hay sự tăng trưởng chậm nhưng ổn định. Chúng ta sẽ khám phá một loạt các kịch bản như vậy trong chương 4 và nghiên cứu chuỗi các hệ quả có khả năng xảy ra trong

chương 5, từ thảm cảnh tới xã hội hoàn hảo không tưởng – vốn đều là các hệ quả gây tranh cãi. Ai sẽ lãnh đạo – con người, AI hay cyborg? Con người sẽ được đối xử tử tế hay tệ hại? Liệu chúng ta có bị thay thế và, nếu có, chúng ta sẽ nhìn nhận người thay thế chúng ta là những kẻ chinh phục hay hậu duệ xứng đáng? Tôi rất tò mò về kịch bản bạn thích trong chương 5! Tôi cũng lập ra trang web <http://AgeOfAI.org> để bạn có thể chia sẻ quan điểm của mình và tham gia cuộc đối thoại.

Cuối cùng, chúng ta sẽ nhồi nhét hàng tỷ năm tương lai vào chương 6, mà trớ trêu thay, ở đó chúng ta có thể rút ra những kết luận đánh thép hơn so với các chương trước, rằng giới hạn tối thượng của sự sống trong vũ trụ chúng ta không phải do trí tuệ mà do các định luật vật lý đặt ra.



Hình 1.6: Câu hỏi nào về AI thú vị hay không sẽ phụ thuộc việc AI nâng cao đến mức độ nào và tương lai của chúng ta sẽ phát triển theo hướng nào.

Sau khi thăm dò về lịch sử của trí tuệ, chúng ta sẽ dành phần còn lại của cuốn sách để xem xét nên hướng tới tương lai như thế nào và cách thức đạt được điều đó. Để có thể liên kết các dữ liệu không khan với các câu hỏi về mục đích và ý nghĩa, chúng ta sẽ khám phá cơ sở vật lý của mục tiêu trong chương 7 và ý thức trong chương 8. Cuối cùng, trong phần kết, chúng ta khám phá những gì có thể thực hiện ngay bây giờ để giúp tạo ra tương lai chúng ta muốn.

	Tiêu đề ngắn của chương	Chủ đề	Trạng thái
Lịch sử của trí tuệ	Phần mở đầu: Câu chuyện về nhóm Omega	Điều đáng suy ngẫm	Hoàn toàn suy đoán
	1 Cuộc đối thoại	Ý tưởng, thuật ngữ chính	Không nhiều suy đoán
	2 Vật chất có được trí tuệ	Bản chất của trí tuệ	
	3 AI, Kinh tế, Vũ khí & Luật pháp	Tương lai gần	
	4 Bùng nổ Trí tuệ	Kịch bản siêu trí tuệ	
	5 Hệ quả	10.000 năm sau	Hoàn toàn suy đoán
Lịch sử của ý nghĩa	6 Di sản của Vũ trụ	Hàng tỷ năm sau	Không nhiều suy đoán
	7 Mục tiêu	Lịch sử của hành vi có mục tiêu	
	8 Ý thức	Ý thức tự nhiên & nhân tạo	
	Hội kết: Câu chuyện về nhóm FLI	Chúng ta nên làm gì?	Không nhiều suy đoán

Hình 1.7: Cấu trúc cuốn sách

Nếu bạn thích đọc nhảy cóc, hầu hết các chương sẽ khá độc lập nếu bạn đã thấm nhuần hết thuật ngữ và định nghĩa trong chương 1 này và phần đầu chương tiếp theo. Nếu bạn là một nhà nghiên cứu AI, bạn có thể chọn bỏ qua toàn bộ chương 2 ngoại trừ phần định nghĩa ban đầu về trí tuệ. Nếu bạn chưa biết về AI, chương 2 và 3 sẽ cung cấp các luận điểm để củng cố cho lý do rằng chương 4 tới chương 6 không nên bị coi nhẹ và xem như khoa học viễn tưởng viễn vông. Hình 1.7 tổng kết tình chất của các chương khác nhau thế nào trên một phổ từ thực tế tới suy đoán.

Một cuộc hành trình hấp dẫn đang đợi chúng ta đấy. Hãy bắt đầu thôi!

TÓM TẮT:

- *Sự sống, được định nghĩa như một quá trình có thể giữ lại sự phức tạp và tái tạo, phát triển qua ba giai đoạn: giai đoạn sinh học (1.0), với phần cứng và phần mềm đều tiến hóa; giai đoạn văn hóa (2.0), khi mà sự sống có thể thiết kế phần mềm của nó (thông qua*

học tập) và giai đoạn công nghệ (3.0), khi nó thiết kế được cả phần cứng, tự làm chủ số phận của chính mình.

- Trí tuệ nhân tạo có thể giúp chúng ta khởi động Sự sống 3.0 vào thế kỷ này, và một cuộc hội thoại thú vị về việc chúng ta nên hướng đến tương lai nào và làm sao để đạt được điều đó đã nảy sinh. Có ba nhóm quan điểm chính trong cuộc tranh luận: phái hoài nghi về công nghệ, phái lý tưởng cách mạng số và phong trào AI-có-ích.

- Những người hoài nghi về công nghệ coi việc xây dựng AGI siêu phàm khó khăn đến mức hàng trăm năm nữa cũng chưa thể thực hiện, nên cho rằng lo lắng về nó (và Sự sống 3.0) từ bây giờ là ngớ ngẩn.

- Phái lý tưởng cách mạng số thấy việc đó có khả năng xảy ra trong thế kỷ này và hết lòng hoan nghênh Sự sống 3.0, coi đây là bước tiếp theo đầy tự nhiên và được mong đợi trong bước tiến hóa vũ trụ.

- Phong trào AI-có-ích cũng thấy nó có khả năng xảy ra trong thế kỷ này, nhưng cho rằng không thể đảm bảo kết quả tốt mà không thực hiện các nhiệm vụ khó khăn như nghiên cứu an toàn AI trước.

- Ngoài những tranh cãi chính đáng giữa các chuyên gia hàng đầu thế giới này, cũng có những tranh cãi giả tạo nhằm chán do hiểu lầm. Ví dụ, đừng bao giờ lãng phí thời gian tranh luận về “sự sống”, “trí tuệ”, hay “ý thức” trước khi chắc chắn rằng bạn và đối tượng cùng tranh luận đang sử dụng những từ này với cùng một định nghĩa! Cuốn sách này sử dụng các định nghĩa trong bảng 1.1.

- Cũng hãy cẩn thận trước những quan niệm sai lầm phổ biến trong hình 1.5: “Siêu trí tuệ vào năm 2100 là không thể tránh khỏi/ không thể xảy ra.” “Chỉ có phe bàn lùi mới lo lắng về AI.” “Chúng ta cần bận tâm về việc AI trở nên ác độc và/hoặc có ý thức, và chỉ còn vài năm nữa thôi.” “Người máy là mối bận tâm chính.” “AI không thể kiểm soát con người và không thể có mục tiêu.”

- Từ chương 2 đến chương 6, chúng ta sẽ khám phá câu chuyện về trí tuệ - từ sự khởi đầu khiêm nhường hàng tỷ năm trước tới các khả năng trong vũ trụ hàng tỷ năm sau này. Trước tiên, chúng ta sẽ đánh giá các thách thức gần - như công việc, vũ khí AI và công cuộc

xây dựng AGI trình độ con người, sau đó khám phá xác suất của những viễn cảnh tương lai thú vị và khả thi cho máy móc và/hoặc con người thông minh. Tôi tự hỏi bạn sẽ thích lựa chọn nào hơn!

- Từ chương 7 đến chương 9, chúng ta sẽ chuyển từ mô tả thực tế khô khan sang khám phá về các mục tiêu, ý thức và ý nghĩa, cũng như nghiên cứu những gì chúng ta có thể làm ngay bây giờ để giúp tạo ra tương lai chúng ta muốn.*

- Tôi cho rằng cuộc hội thoại về tương lai của sự sống có AI là việc quan trọng nhất trong thời đại của chúng ta — hãy cùng tham gia!*

Chương 2

Vật chất có được trí tuệ

Khí hiđro... khi trải qua đủ thời gian, cũng trở thành con người.

Edward Robert Harrison, 1995

Một trong những bước phát triển lớn nhất trong suốt 13,8 tỉ năm kể từ Vụ nổ Lớn là việc vật chất vô tri vô giác đã trở thành những thể sống thông minh. Điều này đã xảy ra thế nào, và trong tương lai các thể sống còn có thể thông minh đến mức nào? Khoa học nói lên điều gì về lịch sử và số phận của trí tuệ trong vũ trụ của chúng ta? Để giúp tìm hiểu những vấn đề này, hãy đọc chương này để khám phá những nền tảng và cơ sở xây dựng nên trí tuệ. Khi nói rằng một khối vật chất có trí tuệ, thì nó phải ra sao? Khi nói rằng một vật thể có thể ghi nhớ, tính toán và học tập thì nó phải ra sao?

Trí tuệ là gì?

Vợ chồng tôi gần đây có may mắn được tham dự một hội nghị chuyên đề về trí tuệ nhân tạo do Quỹ Nobel Thụy Điển tổ chức. Khi một nhóm cử tọa gồm các nhà nghiên cứu AI hàng đầu được đề nghị định nghĩa trí tuệ, họ đã thảo luận rất lâu mà không đi đến thống nhất. Chúng tôi thấy khá buồn cười: ngay cả những nhà nghiên cứu trí tuệ đầy thông minh cũng không nhất trí về định nghĩa trí tuệ! Vậy là rõ ràng không có định nghĩa “chuẩn” nào về trí tuệ cả. Thay vào đó là nhiều định nghĩa cùng tồn tại cạnh tranh lẫn nhau, bao gồm khả năng tư duy logic, hiểu, lên kế hoạch, nhận biết cảm xúc, tự ý thức, sáng tạo, giải quyết vấn đề và học tập.

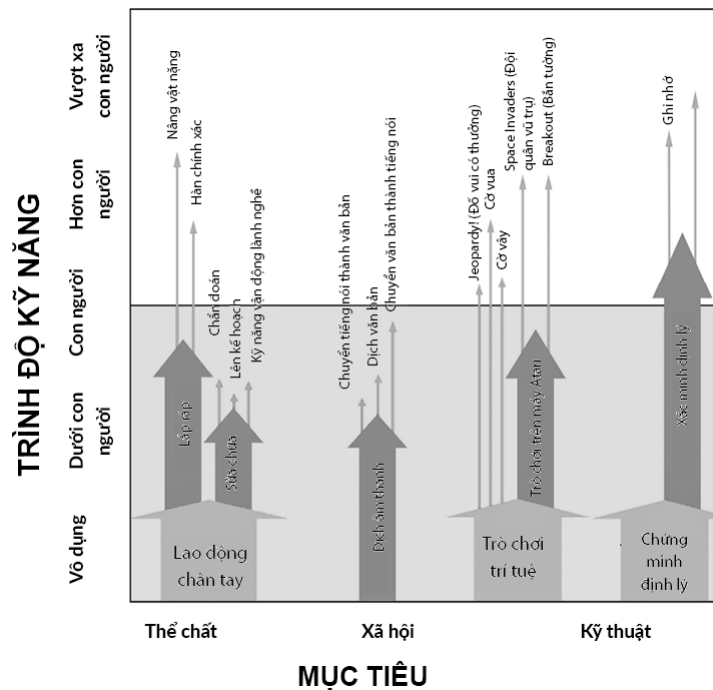
Trong quá trình khám phá tương lai của trí tuệ, chúng ta muốn có được tầm nhìn rộng bao quát tối đa, không giới hạn những loại trí tuệ đã tồn tại từ trước. Vậy nên định nghĩa tôi đưa ra ở chương trước, và cách tôi sẽ dùng từ “trí tuệ” trong suốt cuốn sách này, hết sức rộng:

Trí tuệ = khả năng hoàn thành những mục tiêu phức tạp

Định nghĩa này đủ rộng để bao gồm tất cả các định nghĩa đã nhắc đến ở trên, vì việc hiểu, tự ý thức, giải quyết vấn đề, học tập, v.v. đều là ví dụ của những mục tiêu phức tạp mà người ta có thể có. Định nghĩa này cũng đủ rộng để gói gọn định nghĩa trong Từ điển Oxford—“khả năng thu nhận và ứng dụng kiến thức và kỹ năng”—vì một người có thể coi việc ứng dụng kiến thức và kỹ năng là một mục tiêu.

Vì có thể có rất nhiều mục tiêu khác nhau, nên cũng có thể có rất nhiều loại trí tuệ khác nhau. Bởi vậy, theo định nghĩa của chúng tôi, sẽ vô nghĩa khi đo đếm trí tuệ của con người, động vật không phải người và máy móc chỉ bằng một con số như chỉ số IQ.¹⁶ Cái gì thông minh hơn: một chương trình máy tính chỉ có thể chơi cờ vua hay một chương trình chỉ có thể chơi cờ vây? Không có câu trả lời nào hợp lý cho câu hỏi này, vì chúng giỏi những việc khác nhau, không thể trực tiếp so sánh với nhau. Tuy nhiên, chúng ta có thể nói rằng một chương trình thứ ba thông minh hơn cả hai chương trình kia, nếu ít nhất nó giỏi ngang chúng trong việc hoàn thành tất cả các mục tiêu và giỏi hơn ở ít nhất một mục tiêu (thắng trong trò cờ vua chẳng hạn).

Cãi cọ về việc liệu một người hay vật nào đó có thông minh hay không cũng vô nghĩa, vì khả năng thực hiện điều gì sẽ nằm trên một phổ rộng, chứ không hẳn là một tính chất chỉ bao gồm có hay không. Người nào có khả năng hoàn thành mục tiêu biết nói? Trẻ sơ sinh? Không. Người dẫn chương trình radio? Có. Thế còn trẻ chập chững biết đi và có thể nói được 10 từ thì sao? Hay 500 từ? Bạn sẽ vạch ra ranh giới ở đâu? Tôi đã cố ý dùng một từ mập mờ là “phức tạp” trong định nghĩa trên, vì không có gì thú vị lắm khi ta cố vạch ra một ranh giới nhân tạo giữa trí tuệ và phi-trí tuệ, và việc đo đếm khả năng hoàn thành các mục tiêu khác nhau sẽ có ích hơn.



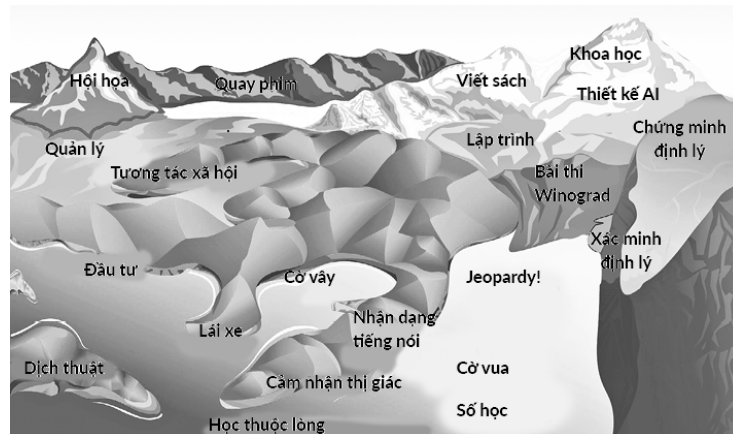
Hình 2.1: Trí tuệ, được định nghĩa là khả năng hoàn thành các mục tiêu phức tạp, không thể đo đếm bằng một chỉ số IQ đơn lẻ, mà chỉ có thể đo bằng một phổ khả năng bao trùm tất cả các mục tiêu. Mỗi mũi tên trên đây chỉ ra mức độ hoàn thiện của các hệ thống AI tốt nhất hiện nay cho việc hoàn thành các mục tiêu khác nhau, cho thấy trí tuệ nhân tạo ngày nay thường ở mức độ hẹp, mỗi hệ thống chỉ có thể hoàn thành những mục tiêu hết sức cụ thể. Ngược lại, trí tuệ con người cực kỳ rộng: một đứa trẻ khỏe mạnh có thể học cách giỏi lên ở hầu hết mọi việc.

Để phân loại trí tuệ một cách khoa học, ta cần phân biệt rõ ràng trí tuệ hẹp và rộng. Thứ đã hạ gục nhà vô địch cờ vua Garry Kasparov năm 1997 là máy tính chơi cờ vua Deep Blue của IBM chỉ có thể hoàn thành một nhiệm vụ rất hẹp là chơi cờ vua— tuy có phần cứng và phần mềm ấn tượng như thế, nhưng nó không thể thắng trong trò cờ caro 9 ô dù chỉ đấu với đứa trẻ bốn tuổi. Hệ thống AI DQN của Google DeepMind có thể hoàn thành nhiều mục tiêu hơn một chút: nó có thể chơi hàng tá trò chơi điện tử Atari cổ điển ngang trình con người hay cao hơn. Ngược lại, trí tuệ con người từ trước đến nay rộng bao la, có khả năng làm chủ một kho tàng kỹ năng đồ sộ. Một đứa trẻ khỏe mạnh, được đào tạo đủ thời gian, có thể giỏi không chỉ

bất kỳ một trò chơi nào, mà còn bất kỳ một ngôn ngữ, môn thể thao hay tay nghề nào. Khi so sánh trí tuệ con người với máy móc hiện nay, loài người chúng ta thắng đẹp về độ rộng của trí tuệ, trong khi máy móc lại hơn ta ở từng miền kỹ năng hẹp, và số miền kỹ năng này tuy còn nhỏ nhưng đang tăng dần, như minh họa trong hình 2.1. Mục tiêu tối cao trong nghiên cứu AI là việc xây dựng được “AI tổng quát” (thường gọi là trí tuệ nhân tạo tổng quát, viết tắt là AGI) rộng tối đa: nó cần có khả năng hoàn thành gần như bất kỳ mục tiêu nào, kể cả học tập. Chúng ta sẽ khám phá chuyện này chi tiết hơn trong chương 4. Thuật ngữ “AGI”, được các nhà nghiên cứu AI là Shane Legg, Mark Gubrud và Ben Goertzel phổ biến rộng rãi có nghĩa cụ thể hơn là trí tuệ nhân tạo tổng quát ở trình độ con người: tức là có khả năng hoàn thành bất kỳ mục tiêu nào ở mức ít nhất cũng giỏi như con người.¹ Tôi sẽ dùng theo định nghĩa của họ, nên trừ khi tôi có ghi rõ kèm theo thuật ngữ viết tắt (ví dụ ghi là “AGI siêu phàm”), thì tôi sẽ dùng “AGI” để nói ngắn gọn cho “AGI ở trình độ con người”.¹⁷

Mặc dù từ “trí tuệ” thường có sắc thái nghĩa tích cực, cũng cần nhấn mạnh rằng chúng ta đang dùng từ này theo cách trung lập hết sức, để chỉ khả năng hoàn thành các mục tiêu phức tạp, bất kể các mục tiêu này được coi là tốt hay xấu. Vì vậy một người thông minh có thể rất giỏi giúp người hoặc hại người. Chúng ta sẽ tìm hiểu vấn đề các loại mục tiêu trong chương 7. Nói về mục tiêu, chúng ta cũng cần yếu tố dù nhỏ nhưng quan trọng rằng chúng ta đang nói đến mục tiêu của ai. Giả sử người máy trợ lý cá nhân mới cứng của bạn trong tương lai không có mục tiêu nào của riêng mình, nhưng sẽ làm bất kỳ điều gì bạn bảo nó làm, và bạn bảo nó nấu một bữa ăn kiểu Ý hoàn hảo. Nếu nó lên mạng và nghiên cứu các công thức nấu bữa tối kiểu Ý, cách để đến siêu thị gần nhất, cách để mì Ý ráo nước và tương tự như vậy, và rồi nó mua được các nguyên liệu và chuẩn bị một bữa ăn ngon lành, có lẽ bạn sẽ coi nó là thông minh, mặc dù mục tiêu ban đầu là của bạn đặt ra. Thực ra, nó đã coi mục tiêu của bạn là của nó một khi bạn đưa ra yêu cầu, và rồi chia nhỏ mục tiêu đó ra thành các cấp độ mục tiêu phụ của riêng nó, từ việc trả tiền cho thu ngân đến

việc bào nhỏ phô mai Parmesan. Theo cách nhìn này, hành vi thông minh không thể không gắn liền với việc đạt được mục tiêu.



Hình 2.2. Tranh minh họa cho “phong cảnh năng lực con người” của Hans Moravec, trong đó các vùng đất nhô cao tượng trưng cho mức độ khó đối với máy tính, còn phần nước biển đang dâng lên tượng trưng cho những gì máy tính hiện có thể làm được.

Chúng ta thường đánh giá độ khó của một nhiệm vụ theo cảm nhận của con người chúng ta khi thực hiện, như trong hình 2.1. Nhưng điều này có thể dẫn đến cái nhìn sai lệch về độ khó của nhiệm vụ đối với máy tính. Ta cảm thấy việc nhân 314159 với 271828 khó hơn việc nhận ra bạn mình trong ảnh, nhưng máy tính đã vượt xa ta trong môn số học từ rất lâu trước khi tôi sinh ra, trong khi việc nhận dạng hình ảnh ở trình độ con người thì máy tính gần đây mới thực hiện được. Việc các nhiệm vụ phối hợp giác quan vận động bậc thấp có vẻ dễ dàng với chúng ta, dù cần đến tài nguyên máy tính khổng lồ được gọi là nghịch lý Moravec; nghịch lý này giải thích rằng chúng ta cảm thấy các nhiệm vụ kể trên thật dễ dàng là do bộ não đã dành riêng cho chúng một lượng khổng lồ phần cứng—trên thực tế là hơn một phần tư não bộ.

Tôi rất thích phép ẩn dụ này của Hans Moravec và đã mạn phép minh họa nó trong hình 2.2:

Máy tính là những cỗ máy mang tính phổ quát; tiềm năng của chúng mở rộng đồng bộ trong vô số nhiệm vụ. Trái lại, tiềm năng của con người mạnh mẽ trong những lĩnh vực từ lâu vốn quan trọng đối với sinh tồn, nhưng lại yếu kém ở những thứ xa rời bản năng ấy. Hãy

tưởng tượng một bức tranh “phong cảnh năng lực con người” có những vùng trũng thấp như “số học” và “học thuộc lòng”, vùng trung du như “chứng minh định lý” và “chơi cờ vua” và những đỉnh núi cao như “vận động”, “phối hợp tay-mắt” và “tương tác xã hội”. Năng lực máy tính đang dần cải tiến giống như nước dâng từ từ tràn ngập phong cảnh này. Nửa thế kỷ trước, nước bắt đầu làm ngập vùng trũng thấp, xóa sổ những người làm công việc tính toán và ghi sổ liệu, nhưng phần lớn chúng ta vẫn còn khô ráo. Giờ cơn lũ đã lan đến vùng trung du và những tiền đồn của chúng ta ở đó đang tính tới chuyện thoái lui. Chúng ta cảm thấy vẫn an toàn trên những đỉnh núi cao, nhưng với tốc độ hiện nay, ngay cả những đỉnh núi đó cũng sẽ ngập trong nước chỉ trong vòng nửa thế kỷ nữa. Tôi xin đề xuất rằng chúng ta nên lắp những con thuyền lớn như thuyền của Noah khi ngày đó tới gần và tập sống chung với lũ!²

Trong các thập kỷ sau khi ông viết những dòng này, mực nước biển ấy vẫn tiếp tục dâng không ngừng nghỉ, như ông dự đoán, giống như hiện tượng nóng lên toàn cầu hoạt động điên cuồng, và một số vùng trung du (bao gồm cả lĩnh vực cờ vua) từ lâu đã chìm trong nước. Điều xảy ra tiếp đến và điều chúng ta nên làm chính là chủ đề phần còn lại của cuốn sách này.

Nếu nước biển này tiếp tục dâng, một ngày nào đó nó có thể chạm tới điểm bùng phát, châm ngòi cho một thay đổi to lớn. Mực nước biển hệ trọng này tương ứng với thời điểm máy móc có khả năng thực hiện nhiệm vụ thiết kế AI. Trước khi đến điểm bùng phát đó, nước biển dâng lên là do con người cải tiến máy móc; sau đó, nước biển sẽ dâng do máy móc cải tiến máy móc, có khả năng nhanh hơn rất nhiều so với những gì con người có thể làm và sẽ nhanh chóng làm ngập tất cả đất đai. Đây chính là ý tưởng vừa lý thú vừa gây tranh cãi về điểm kỳ dị mà chúng ta sẽ cùng khám phá trong chương 4.

Nhà tiên phong về máy tính Alan Turing từng có chứng minh nổi tiếng rằng nếu một máy tính có thể thực hiện một tập lệnh tối thiểu, thì với đủ thời gian và bộ nhớ, nó có thể được lập trình để thực hiện điều mà bất kỳ máy tính nào có thể làm được. Máy móc vượt qua ngưỡng trọng yếu này sẽ được gọi là máy tính phổ quát (hay máy

tính phổ quát Turing); tất cả điện thoại thông minh và máy tính cá nhân ngày nay đều mang tính phổ quát theo nghĩa này. Tương tự, tôi cho rằng ngưỡng trí tuệ trọng yếu cần thiết cho thiết kế AI là một ngưỡng trí tuệ phổ quát: nếu có đủ thời gian và tài nguyên, trí tuệ đó có thể tự khiến mình có khả năng hoàn thành bất kỳ mục tiêu nào với trình độ tương đương bất kỳ thực thể thông minh nào. Ví dụ, nếu nó quyết định nó muốn giỏi hơn về kỹ năng xã hội, kỹ năng dự đoán hay kỹ năng thiết kế AI, nó có thể đạt được điều đó. Nếu nó quyết định tìm ra cách để xây dựng một nhà máy rô-bốt, nó có thể làm vậy. Nói cách khác, trí tuệ phổ quát có tiềm năng phát triển thành Sự sống 3.0.

Quan điểm thường thấy trong cộng đồng nghiên cứu về trí tuệ nhân tạo là trí tuệ xét cho cùng được tạo thành bởi thông tin và sự tính toán, không phải là máu thịt hay các nguyên tử cac-bon. Điều này đồng nghĩa rằng sẽ không có lý do nào khiến máy móc không thể một ngày nào đó thông minh ít nhất bằng chúng ta.

Nhưng thật sự thì thông tin và tính toán là gì, khi mà vật lý học đã dạy chúng ta rằng ở mức độ cơ bản, mọi thứ chỉ là vật chất và năng lượng đang chuyển động mà thôi? Làm thế nào một thứ trừu tượng, phi vật thể và hư vô như thông tin và tính toán có thể mang hình hài những vật thể chạm vào được? Cụ thể, làm thế nào một mớ hạt vô tri chuyển động theo các định luật vật lý có thể thể hiện những hành vi mà ta gọi là thông minh?

Nếu bạn thấy câu trả lời cho câu hỏi này đã rõ ràng và coi việc máy móc thông minh như con người là có thể xảy ra trong thế kỷ này —vì bạn là một nhà nghiên cứu AI chẳng hạn—thì xin vui lòng bỏ qua phần còn lại của chương này và nhảy thẳng đến chương 3. Nếu không, bạn sẽ hài lòng khi biết rằng tôi đã viết hẳn ba mục tiếp theo dành riêng cho bạn.

Trí nhớ là gì?

Nếu ta nói một cuốn địa đồ chứa đựng

thông tin

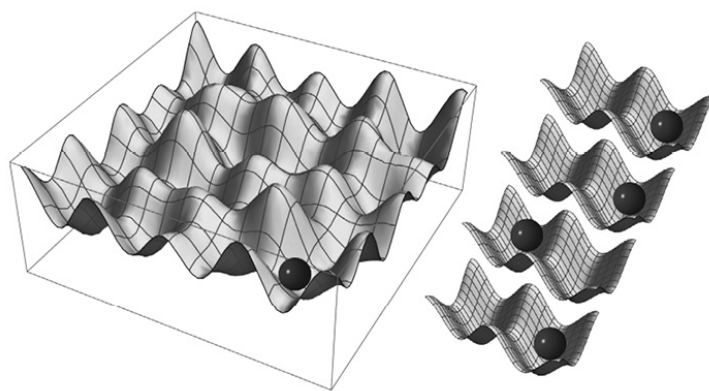
về thế giới, tức là có một mối liên quan giữa trạng thái của cuốn sách (cụ thể là vị trí của các phân tử nhất định có chức năng đem lại màu sắc trên các chữ cái và hình ảnh) với trạng thái của thế giới (ví dụ, vị trí của các lục địa). Nếu các lục địa thay đổi vị trí thì các phân tử trên trang sách cũng phải ở vị trí khác. Con người sử dụng một loạt các thiết bị để lưu trữ thông tin, từ sách vở và não bộ đến ổ cứng, và chúng đều có chung một đặc điểm: trạng thái của chúng có thể có liên quan tới (và vì vậy đem lại thông tin cho chúng ta về) trạng thái của những thứ khác mà ta quan tâm.

Đặc điểm vật lý cơ bản chung nào giữa những thiết bị này khiến chúng trở thành những công cụ ghi nhớ hữu ích, tức những thiết bị để lưu trữ thông tin? Câu trả lời là chúng đều có thể thể hiện các trạng thái kéo dài khác nhau—đủ để mã hóa thông tin cho đến khi cần đến thông tin đó. Giả sử bạn đặt một quả bóng trên một địa hình dạng đồi núi với 16 thung lũng khác nhau, như trong hình 2.3. Một khi quả bóng đã lăn xuống và dừng lại, nó sẽ nằm tại một trong 16 vị trí; như vậy, bạn có thể dùng vị trí của nó để nhớ một số bất kỳ trong khoảng từ 1 đến 16.

Công cụ ghi nhớ này khá chắc chắn, vì ngay cả khi bị ngoại lực tác động thì nhiều khả năng quả bóng vẫn sẽ nằm tại thung lũng ban đầu, nên bạn vẫn sẽ chỉ ra được con số đang được ghi nhớ. Ký ức này bền như vậy bởi nếu muốn đưa quả bóng ra khỏi thung lũng, chúng ta sẽ cần nhiều năng lượng hơn là chỉ tác động các lực ngẫu nhiên. Ý tưởng này cũng có thể giải thích cho những ký ức bền nói chung, không chỉ là một quả bóng chuyển động: năng lượng của một hệ thống vật lý phức tạp có thể phụ thuộc vào đủ loại đặc điểm cơ học, hóa học, điện và từ, và chừng nào việc thay đổi trạng thái của hệ thống so với trạng thái mà bạn muốn nó ghi nhớ vẫn cần đến năng lượng, thì trạng thái đó vẫn bền vững. Thế nên, chất rắn có rất nhiều trạng thái kéo dài khác nhau, còn chất lỏng và chất khí thì không: nếu bạn khắc tên ai đó lên một chiếc nhẫn vàng, thông tin đó vẫn sẽ tồn tại nhiều năm sau vì chúng ta sẽ cần nhiều năng lượng để đúc lại kim loại vàng, nhưng nếu bạn viết tên người đó lên mặt nước hồ thì nó sẽ biến mất trong tích tắc vì mặt nước rất dễ thay hình đổi dạng.

Công cụ ghi nhớ đơn giản nhất chỉ có hai trạng thái bền vững (hình 2.3, trang x). Do đó, chúng ta có thể coi nó như việc mã hóa một

số nhị phân (binary digit, viết tắt là “bit”), tức là số 0 hoặc số 1. Thông tin chứa trong các công cụ ghi nhớ phức tạp hơn sẽ được chứa trong nhiều bit hơn: ví dụ, 4 bit trong hình 2.3 (bên phải) gộp lại có thể nằm trong $2 \times 2 \times 2 \times 2 = 16$ trạng thái khác nhau: 0000, 0001, 0010, 0011, ... , 1111. Nhờ vậy, tổ hợp này có khả năng ghi nhớ tương đương với một hệ thống 16 trạng thái phức tạp hơn (bên trái). Vì thế, chúng ta có thể coi các bit như những nguyên tử thông tin—khối thông tin nhỏ nhất có thể phân chia được, không thể chia nhỏ hơn nữa và có thể kết hợp để tạo thành bất kỳ thông tin nào. Ví dụ, khi tôi gõ từ “word”, máy tính sẽ biểu thị từ này trong bộ nhớ bằng một dãy gồm 4 số, 119 111 114 100, và lưu trữ mỗi số đó bằng 8 bit dữ liệu (mỗi chữ cái không viết hoa được biểu đạt bằng một số có giá trị bằng 96 cộng với số thứ tự của chữ cái đó trong bảng chữ cái). Ngay khi tôi gõ phím w trên bàn phím, máy tính sẽ hiển thị hình ảnh chữ cái w trên màn hình, và hình ảnh này cũng được biểu thị bởi các bit: 32 bit sẽ xác định màu sắc của mỗi điểm ảnh trong hàng triệu điểm ảnh trên màn hình.



Hình 2.3: Một vật thể vật lý sẽ có khả năng ghi nhớ hữu ích nếu nó có thể tồn tại ở nhiều trạng thái bền vững khác nhau. Quả bóng bên trái có thể mã hóa 4 bit thông tin cho biết nó đang ở đâu trong $2^4 = 16$ thung lũng. Bốn quả bóng bên phải khi kết hợp lại cũng có thể mã hóa 4 bit thông tin—mỗi quả một bit.

Vì hệ thống hai trạng thái rất dễ sản xuất và điều khiển nên hầu hết máy tính hiện đại đều lưu trữ thông tin dưới dạng bit, nhưng các bit này được thể hiện dưới muôn hình vạn trạng. Trên một đĩa DVD,

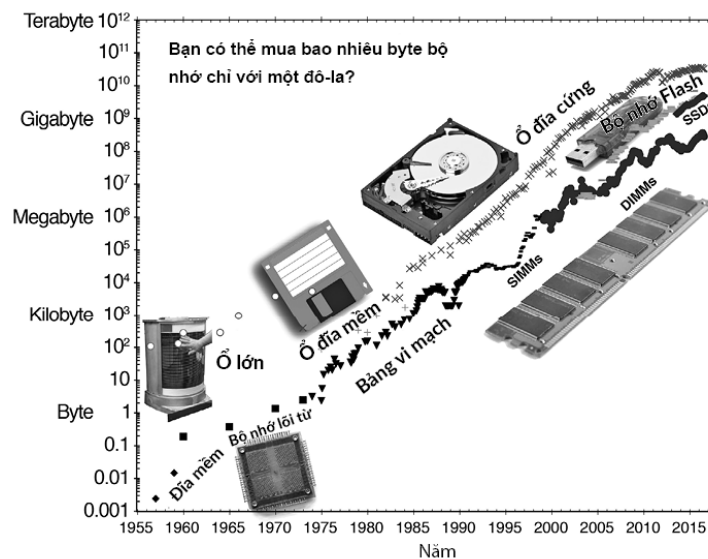
mỗi bit sẽ tương ứng với một rãnh siêu vi có hay không xuất hiện tại một điểm nhất định trên bề mặt đĩa. Trên ổ cứng, mỗi bit tương ứng với một trong hai trạng thái nhiễm từ trên một điểm trên bề mặt đĩa. Trong bộ nhớ máy tính, mỗi bit tương ứng với vị trí của những electron nhất định, xác định xem một vi tụ điện có mang điện hay không. Một số loại bit cũng có thể dễ dàng truyền tải, kể cả với tốc độ ánh sáng: ví dụ, trong một cáp quang gửi thư điện tử (email), mỗi bit tương ứng với một tia sáng laser mạnh hay yếu ở từng thời điểm.

Các kỹ sư thường mã hóa bit vào những hệ thống không những bền vững và dễ đọc (như một chiếc nhẫn vàng), mà còn dễ viết nữa. Để thay đổi trạng thái một ổ cứng, ta cần ít năng lượng hơn nhiều so với việc khắc lên vàng. Họ cũng hay dùng những hệ thống tiện sử dụng và có chi phí sản xuất hàng loạt rẻ. Nhưng ngoài việc đó, họ sẽ không quan tâm việc các bit sẽ có dạng vật lý nào—và bạn cũng sẽ vậy phần nhiều thời gian, vì chuyện đó chẳng quan trọng! Nếu bạn gửi email một văn bản cho một người bạn để in ra, thông tin có thể được sao chép nhanh chóng liên tục từ dạng các khu vực từ hóa trong ổ cứng thành các điện tích trong bộ nhớ máy tính, sóng radio của mạng không dây, điện áp trong bộ định tuyến (router), xung laser trong cáp quang, và cuối cùng là các phân tử trên tờ giấy. Nói cách khác, thông tin có thể có sự sống riêng, độc lập với bản chất vật lý của nó! Quả thực, thường chúng ta chỉ quan tâm đến khía cạnh độc lập với bản chất vật lý của thông tin: nếu người bạn kia gọi điện thoại cho bạn để bàn về văn bản bạn vừa gửi, chắc chắn cô sẽ không nói đến điện áp hay phân tử. Đây là điều đầu tiên gợi cho chúng ta biết làm thế nào một thứ phi vật thể như trí tuệ có thể được thể hiện thành những vật thể, và chúng ta sẽ sớm thấy được chiều sâu của ý tưởng về sự độc lập với bản chất vật lý, bao gồm không chỉ thông tin mà còn cả tính toán và học tập.

Nhờ có sự độc lập với bản chất vật lý này, các kỹ sư khôn ngoan đã có thể liên tục thay thế những thiết bị ghi nhớ trong máy tính bằng những thiết bị tốt hơn rất nhiều, dựa vào công nghệ mới, mà không cần thay đổi phần mềm chút nào. Kết quả thật đáng ngưỡng mộ, như đã minh họa trong hình 2.4: trong sáu thập kỷ qua, cứ đôi ba năm bộ nhớ máy tính lại giảm một nửa giá thành. Ổ cứng đã giảm giá thành hơn 100 triệu lần và bộ nhớ tốc độ cao hơn dùng cho tính toán thay vì

lưu trữ thì đã rẻ đi đến 10 ngàn tỉ lần. Nếu tất cả những gì bạn mua được “giảm giá 99.9999999999%” như vậy, bạn có thể mua được tất cả bất động sản ở thành phố New York với giá 10 xu Mỹ và tất cả vàng đã từng khai thác trên thế giới với giá một đô-la Mỹ.

Đối với hầu hết chúng ta, sự cải tiến vượt bậc trong công nghệ bộ nhớ thường đi liền với những câu chuyện cá nhân. Tôi vẫn vui khi nhớ lại hồi làm việc ở một cửa hàng kẹo thời trung học để kiếm tiền mua một chiếc máy tính có bộ nhớ những 16 kilobyte (kB), rồi khi tôi lập trình và bán một phần mềm xử lý văn bản cho chiếc máy tính đó với người bạn cùng lớp Magnus Bodin, chúng tôi buộc phải viết phần mềm đó bằng mã máy siêu nhỏ gọn để dành đủ bộ nhớ cho văn bản mà nó cần xử lý. Sau khi đã quen với những đĩa mềm chỉ có dung lượng 70kB, tôi đã choáng ngợp trước loại đĩa mềm nhỏ hơn, có 9 cm mà chứa được đến 1,44MB bằng cả cuốn sách, và rồi ổ cứng đầu tiên của tôi có dung lượng 10MB—có khi chỉ vừa đủ cho một bài hát khi tải về ngày nay. Những kỷ niệm thời niên thiếu này thực sự rất ảo diệu, khi mới chỉ hôm trước thôi, tôi chỉ cần bỏ ra 100 đô-la để mua một chiếc ổ cứng với dung lượng gấp 300.000 lần.



Hình 2.4: Trong sáu thập kỷ qua, cứ khoảng hai năm bộ nhớ máy tính lại giảm một nửa giá thành, tương đương với việc giá thành giảm đi hàng ngàn lần sau mỗi 20 năm. Một byte bằng 8 bit. Dữ liệu

do John McCallum cung cấp, từ nguồn <http://www.jcmit.net/memoryprice.htm>.

Thế còn những công cụ ghi nhớ tự tiến hóa thay vì được con người thiết kế thì sao? Tuy các nhà sinh vật học vẫn chưa biết dạng sống đầu tiên biết sao chép bản vẽ thiết kế chính mình cho thế hệ sau là loài nào, nhưng có vẻ nó từng khá nhỏ. Năm 2016, một nhóm nghiên cứu do Philipp Holliger chỉ đạo ở Đại học Cambridge đã tạo ra một phân tử RNA¹⁸, mã hóa 412 bit thông tin di truyền và có khả năng sao chép các chuỗi RNA dài hơn chính nó; điều này đã củng cố giả thuyết “thế giới RNA” rằng sự sống sơ khai trên Trái Đất bao gồm những chuỗi RNA ngắn tự sao chép. Cho tới nay, công cụ ghi nhớ nhỏ nhất được biết có thể tiến hóa và sử dụng trong tự nhiên là bộ gen của vi khuẩn *Candidatus Carsonella ruddii*, có dung lượng 40 kilobyte, trong khi DNA của loài người chúng ta chứa đựng khoảng 1,6 gigabyte, tương đương với một bộ phim khi tải về. Như đã đề cập trong chương trước, não bộ chúng ta chứa nhiều thông tin hơn trong gen: khoảng 10 gigabyte xét theo điện học (nếu chỉ xét nơ-ron nào trong 100 tỉ nơ-ron đang phát điện tại thời điểm bất kỳ) và 100 terabyte xét theo hóa học/sinh học (nếu xét cả mức độ mạnh yếu của nơ-ron khi kết nối với các xi-nap). So sánh các con số này với bộ nhớ của máy cho thấy những máy tính tốt nhất thế giới hiện nay có thể ghi nhớ nhiều hơn bất kỳ hệ thống sinh học nào—với chi phí ngày càng giảm nhanh chóng, vào năm 2016 chỉ còn khoảng vài ngàn đô-la Mỹ.

Bộ nhớ trong não hoạt động rất khác với bộ nhớ máy tính, không chỉ về cấu tạo mà còn về cách sử dụng. Trong khi bạn có thể lấy dữ liệu lưu trữ trong máy tính hay ổ cứng bằng thông tin nó được lưu ở đâu, thì bạn lấy ký ức từ não bằng thông tin ký ức được lưu đó là gì. Mỗi nhóm bit trong bộ nhớ máy tính có một địa chỉ bằng số và để lấy ra một mẫu thông tin, máy tính sẽ theo thông tin địa chỉ đó để tìm, cũng như khi tôi nói bạn hãy “Đi đến tủ sách của tôi, lấy cuốn thứ năm từ phải sang ở ngăn trên cùng, rồi nói cho tôi biết nội dung trên trang 314”. Ngược lại, bạn lấy thông tin từ não bộ theo cách tương tự như khi bạn lấy thông tin từ một công cụ tìm kiếm: bạn chỉ ra một phần của thông tin hay thứ gì đó liên quan tới nó, rồi nó sẽ hiện ra.

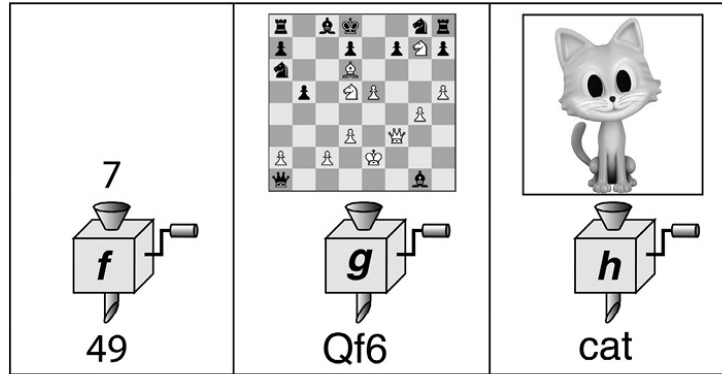
Nếu tôi nói với bạn câu “tồn tại hay không”, hay nếu tôi tra Google câu đó, sẽ có khả năng câu nói “Tồn tại hay không tồn tại, đó là câu hỏi” sẽ hiện ra. Kể cả, nếu tôi có dùng một phần khác trong câu danh ngôn đó hay nhầm lẫn một vài từ thì có lẽ vẫn sẽ tìm được kết quả. Hệ thống ghi nhớ như vậy gọi là tự động liên tưởng, vì chúng nhớ lại thông qua liên tưởng thay vì thông qua địa chỉ.

Trong một luận văn nổi tiếng năm 1982, nhà vật lý học John Hopfield đã cho thấy cách một mạng lưới các nơ-ron kết nối với nhau có thể thực hiện chức năng của bộ nhớ tự động liên tưởng. Tôi thấy ý tưởng cơ bản rất đẹp và có hiệu quả với bất kỳ hệ thống vật lý nào gồm nhiều trạng thái bền vững. Ví dụ, hãy xem xét một quả bóng trên bề mặt gồm hai thung lũng, như hệ thống một-bit trong hình 2.3, và tạo hình cho bề mặt sao cho tọa độ x của hai điểm cực tiểu nơi quả bóng có thể dừng lại lần lượt là $x = \sqrt{2} \approx 1.41421$ và $x = \pi \approx 3.14159$. Nếu chỉ nhớ là π khoảng bằng 3, bạn chỉ cần đặt quả bóng vào tọa độ $x = 3$ và để nó tự lăn tới điểm cực tiểu gần nhất, để một giá trị π chính xác hơn lộ ra. Hopfield nhận ra một hệ thống nơ-ron phức tạp cung cấp một phong cảnh tương tự với rất nhiều điểm cực tiểu năng lượng cho hệ thống lọt vào, và sau đó khoa học đã chứng minh rằng bạn có thể nhồi nhét đến 138 ký ức khác nhau vào mỗi 1000 nơ-ron mà không gây nhầm lẫn đáng kể.

Tính toán là gì?

Giờ ta đã thấy cách một vật thể vật lý ghi nhớ thông tin. Nhưng nó tính toán bằng cách nào?

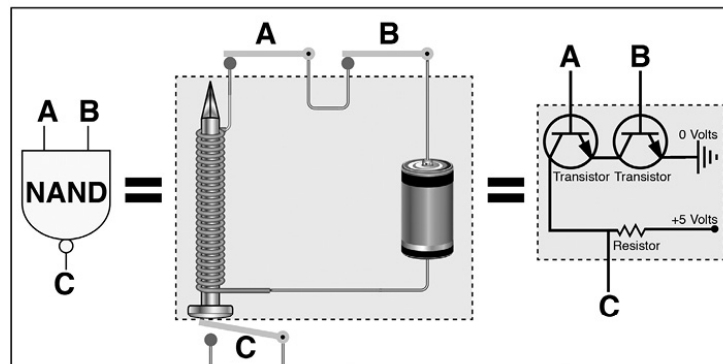
Một phép tính toán là sự biến đổi một trạng thái bộ nhớ sang một trạng thái khác. Nói cách khác, tính toán là việc lấy thông tin và biến đổi nó, thực hiện cái mà các nhà toán học gọi là một hàm số. Tôi coi hàm số giống như một chiếc máy xay cho món thịt thông tin, như minh họa trong hình 2.5: bạn đưa thông tin vào từ trên, quay cần xay và nhận được thông tin đã xử lý chui ra phía dưới—và bạn có thể lặp lại quy trình này bao nhiêu lần cũng được với những thông tin đầu vào khác nhau. Quá trình xử lý thông tin này mang tính tất định, tức là nếu bạn lặp lại quá trình này với cùng thông tin đầu vào thì bạn cũng sẽ có được cùng thông tin đầu ra.



Hình 2.5: Một phép tính toán lấy thông tin và biến đổi nó, thực hiện cái mà các nhà toán học gọi là một hàm số. Hàm số f (bên trái) lấy các bit biểu đạt một số và tính toán bình phương của số đó. Hàm số g (ở giữa) lấy các bit biểu đạt một vị trí quân cờ và tính toán nước đi tốt nhất cho quân trắng. Hàm số h (bên phải) lấy các bit biểu đạt một hình ảnh và tính toán ra nhãn mô tả hình ảnh đó.

Nghe đơn giản là vậy, nhưng khái niệm hàm số này bao quát rộng đến khó tin. Một số hàm số khá đơn giản, như hàm NOT với đầu vào là một bit đơn lẻ và đầu ra là bit trái ngược với nó, khiến 0 biến thành 1 và ngược lại. Các hàm số chúng ta học ở trường phổ thông thường ứng với các nút trên máy tính bỏ túi, đầu vào là một hay nhiều hơn các con số và đầu ra là một con số duy nhất—ví dụ, hàm số $x2$ có đầu vào là một số và đầu ra là số đó tự nhân với chính mình. Các hàm số khác lại cực kỳ phức tạp. Ví dụ, nếu bạn sở hữu một hàm số có đầu vào là các bit đại diện cho các vị trí ngẫu nhiên của quân cờ trên bàn cờ và đầu ra là các bit đại diện cho nước đi tiếp theo tốt nhất, bạn có thể dùng hàm số đó để thắng giải vô địch cờ vua máy tính thế giới. Nếu bạn sở hữu một hàm số có đầu vào là tất cả dữ liệu tài chính trên thế giới và đầu ra là những cổ phiếu tốt nhất nên mua, bạn sẽ nhanh chóng giàu to. Nhiều nhà nghiên cứu AI đã dành cả sự nghiệp để tìm ra cách thực hiện những hàm số nhất định. Chẳng hạn, mục tiêu của nghiên cứu về dịch máy là thực hiện một hàm số có đầu vào là các bit biểu đạt văn bản ở một thứ tiếng và đầu ra là các bit biểu đạt chính văn bản đó nhưng ở thứ tiếng khác, còn mục tiêu của nghiên cứu về tự động mô tả ảnh là đưa vào các bit

biểu đạt một hình ảnh và đưa ra các bit biểu đạt văn bản mô tả hình ảnh đó (hình 2.5, bên phải).



Hình 2.6: Một cổng NAND có hai bit đầu vào, A và B, và tính toán một bit đầu ra là C, theo quy luật $C = 0$ nếu $A = B = 1$ và $C = 1$ trong các trường hợp khác. Ta có thể dùng nhiều hệ thống vật lý như cổng NAND. Trong ví dụ ở giữa, các công tắc tượng trưng cho bit với giá trị 0 = mở, 1 = đóng, và khi công tắc A và B cùng đóng, nam châm điện mở công tắc C. Trong ví dụ bên phải, điện áp (hiệu điện thế) được diễn dịch thành bit khi 1 = 5 volt, 0 = 0 volt, và khi hai dây A và B cùng ở 5 volt, hai transistor dẫn điện và dây C tụt xuống khoảng 0 volt.

Nói cách khác, nếu bạn có thể thực hiện những hàm số vô cùng phức tạp thì bạn cũng có thể phát triển một cỗ máy thông minh có khả năng hoàn thành các mục tiêu vô cùng phức tạp. Điều này khiến câu hỏi về cách vật chất có thể trở nên thông minh đi vào trọng tâm hơn: làm thế nào mà một khối vật chất tưởng như vô tri lại có thể tính toán một hàm số phức tạp?

Thay vì cứ mãi bất động như một chiếc nhẫn vàng hay các công cụ ghi nhớ tĩnh khác, nó phải thể hiện những động thái phức tạp để trạng thái tương lai của nó phụ thuộc vào (và hy vọng là có thể điều khiển/lập trình được) trạng thái hiện tại theo cách lắt léo nào đó. Cách bố trí nguyên tử của nó phải không trật tự như một chất rắn, vốn chẳng có thay đổi gì thú vị, nhưng lại phải trật tự hơn chất lỏng hay chất khí. Cụ thể, chúng ta muốn hệ thống có đặc tính là nếu ta đặt nó vào trạng thái mã hóa thông tin đầu vào, để nó tiến hóa theo định luật vật lý trong một thời gian, rồi diễn dịch trạng thái kết quả

cuối cùng thành thông tin đầu ra, thì thông tin đầu ra đó là hàm số mong muốn khi nhập đầu vào. Nếu điều này thành sự thật thì ta có thể nói là hệ thống tính toán được hàm số.

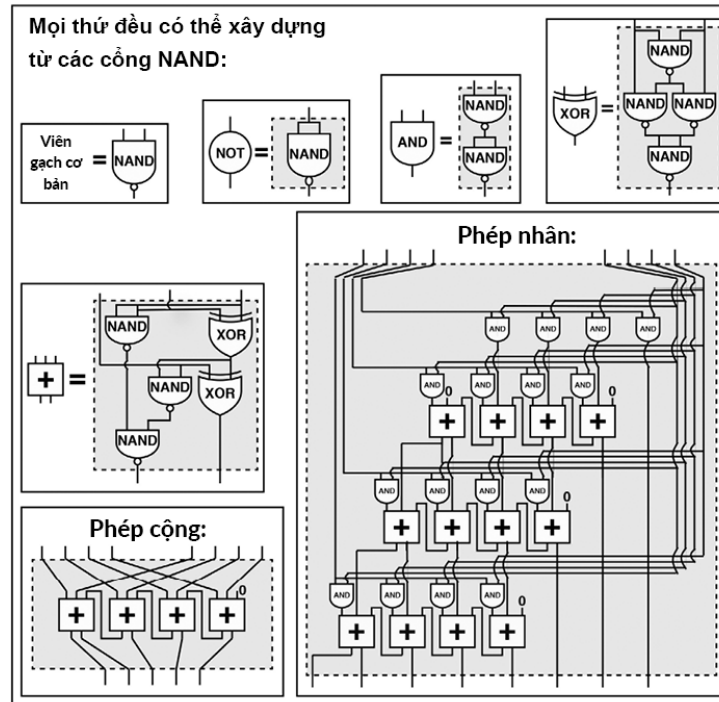
Để có ví dụ đầu tiên cho ý tưởng này, ta hãy thử xây dựng một hàm rất đơn giản (nhưng cũng rất quan trọng) gọi là cổng NAND¹⁹ từ vật chất vô tri. Hàm số này có đầu vào là hai bit và đầu ra là một bit: đầu ra là 0 nếu cả hai đầu vào là 1; với tất cả các trường hợp còn lại thì đầu ra là 1. Nếu ta nối tiếp hai công tắc với một pin và một nam châm điện thì nam châm điện sẽ chỉ bật khi cả công tắc thứ nhất và công tắc thứ hai đều đóng (trạng thái “bật”). Hãy thêm một công tắc thứ ba dưới nam châm điện, như minh họa trong hình 2.6, sao cho nam châm sẽ mở công tắc bất cứ khi nào điện chạy qua. Nếu ta diễn giải hai công tắc đầu là hai bit đầu vào và công tắc thứ ba là bit đầu ra (với 0 = công tắc mở, 1 = công tắc đóng), thì ta đã có một cổng NAND: công tắc thứ ba mở chỉ khi cả hai công tắc đầu đều đóng. Có rất nhiều cách khác để xây dựng các cổng NAND thực tiễn hơn—ví dụ, sử dụng transistor như minh họa trong hình 2.6 (trang X). Trong các máy tính ngày nay, các cổng NAND thường được xây dựng từ các transistor siêu nhỏ và các linh kiện khác vốn được tự động khắc lên các tấm silicon.

Một định lý phi thường trong khoa học máy tính cho rằng các cổng NAND mang tính phổ quát, nghĩa là bạn có thể thực hiện bất kỳ hàm số hoàn thiện nào chỉ bằng cách kết nối các cổng NAND với nhau²⁰. Vậy là nếu có thể xây dựng đủ số cổng NAND, bạn có thể xây dựng một thiết bị tính toán bất kỳ thứ gì! Trong trường hợp bạn muốn thử tìm hiểu cách định lý này hoạt động, tôi đã minh họa trong hình 2.7 cách làm phép nhân chỉ bằng cổng NAND.

Các nhà nghiên cứu của Đại học MIT là Norman Margolus và Tommaso Toffoli đã sáng tạo ra thuật ngữ computronium (vật liệu thông minh) để chỉ bất kỳ chất nào có thể thực hiện tính toán ngẫu nhiên. Chúng ta vừa thấy rằng tạo ra computronium không hẳn quá khó: chất đó chỉ cần có khả năng sử dụng các cổng NAND kết nối với nhau theo cách ta muốn. Quả thực, có vô số các loại computronium khác. Một biến thể đơn giản cũng tạo thành computronium khi thay thế các cổng NAND bằng các cổng NOR cho đầu ra là 1 khi cả hai

đầu vào là 0. Trong mục tiếp theo, chúng ta sẽ tìm hiểu mạng nơ-ron, vốn cũng có khả năng thực hiện tính toán ngẫu nhiên (tức là có vai trò như computronium). Nhà khoa học và doanh nhân Stephen Wolfram cho thấy điều tương tự cũng xảy ra đối với các thiết bị đơn giản gọi là tế bào tự động, là các tế bào liên tục cập nhật các bit dựa trên hoạt động của các bit lân cận. Ngay từ năm 1936, nhà tiên phong về máy tính Alan Turing đã chứng minh trong một luận văn bước ngoặt là một cỗ máy đơn giản (nay gọi là “máy phổ quát Turing”) vừa có thể chi phối các biểu tượng trên một dải băng vừa có thể thực hiện tính toán ngẫu nhiên. Tóm lại, vật chất không chỉ có thể thực hiện bất kỳ phép tính toán hoàn thiện nào mà còn có thể thực hiện theo vô số cách khác nhau.

Như đã đề cập trước đó, Turing cũng chứng minh một điều còn lớn lao hơn trong một luận văn năm 1936: nếu một loại máy tính có thể thực hiện một tập lệnh tối thiểu thì nó mang tính phổ quát, tức nếu có đủ tài nguyên thì nó có thể làm được bất kỳ điều gì các máy tính khác có thể làm. Ông đã cho thấy máy Turing mang tính phổ quát, và khi nối lại với nguồn gốc vật lý, ta lại thấy rằng họ hàng nhà máy tính phổ quát cũng bao gồm những vật thể rất khác nhau, từ một mạng lưới các cổng NAND đến một mạng lưới các nơ-ron kết nối với nhau. Trên thực tế, Stephen Wolfram đã tranh luận rằng hầu hết các hệ thống vật lý không tầm thường, từ một hệ thời tiết đến hệ thần kinh, đều là những máy tính phổ quát nếu chúng đủ lớn và bền.



Hình 2.7: Bất kỳ phép tính toán hoàn thiện nào đều có thể thực hiện bằng cách khéo léo kết hợp chỉ các cổng NAND với nhau. Ví dụ, các mô-đun phép cộng và phép nhân ở trên đều có đầu vào là hai số nhị phân biểu đạt bởi 4 bit và đầu ra là một số nhị phân biểu đạt bởi lần lượt là 5 bit và 8 bit. Các mô-đun nhỏ hơn gồm NOT, AND, XOR và + (cộng ba bit riêng biệt thành một số nhị phân 2 bit) đều từ các cổng NAND tạo thành. Chúng ta khó mà hiểu hết được hình minh họa này, và nếu có hiểu thì chuyện này cũng không cần thiết để giúp bạn hiểu được phần còn lại của cuốn sách; tôi đưa vào đây chỉ để minh họa ý tưởng về tính phổ quát—và để thỏa mãn con người đầu to mắt cận trong tôi.

Việc có thể thực hiện chính xác cùng một phép tính toán trên bất kỳ máy tính phổ quát nào nghĩa là tính toán không phụ thuộc vào bản chất vật lý cũng y như thông tin vậy: nó có thể có sự sống riêng, độc lập với bản chất vật lý của nó! Vì vậy, nếu bạn là một nhân vật siêu thông minh có ý thức trong một trò chơi điện tử trong tương lai, bạn sẽ không có cách nào biết được mình chạy trên máy tính để bàn chạy hệ điều hành Windows, máy tính xách tay chạy hệ điều hành Mac hay điện thoại chạy hệ điều hành Android, bởi bạn độc lập với

bản chất vật lý. Bạn cũng sẽ không có cách nào biết được loại transistor nào đang dùng trong bộ vi xử lý.

Tôi bắt đầu thực sự coi trọng ý tưởng sự vật độc lập với bản chất vật lý này bởi đã chứng kiến nhiều ví dụ tuyệt đẹp trong vật lý học. Chẳng hạn như sóng: sóng có các đặc tính như tốc độ, bước sóng và tần số, và các nhà vật lý chúng tôi có thể nghiên cứu các phương trình sóng mà không cần biết sóng đó nằm trong chất cụ thể nào. Khi bạn nghe thấy tiếng gì đó, bạn đang phát hiện các sóng âm thanh tạo ra bởi các phân tử va đập trong hỗn hợp các chất khí mà ta gọi là không khí, và ta có thể tính toán đủ thú vị về những sóng này—như cường độ của chúng giảm dần theo bình phương của khoảng cách, chúng bẻ cong khi đi qua cửa mở và dội lại khi gặp các bức tường gây ra tiếng vọng—mà chẳng cần biết không khí tạo thành từ những chất gì. Thực ra ta thậm chí chẳng cần biết không khí làm từ các phân tử: ta có thể bỏ qua tất cả các tiểu tiết về oxi, nito, cacbon điôxit, v.v. vì đặc tính duy nhất quan trọng của sóng có trong phương trình sóng nổi tiếng là một con số duy nhất ta có thể đo được: tốc độ sóng, trong trường hợp sóng âm thanh là khoảng 300 mét trên giây. Trên thực tế, phương trình sóng mà tôi dạy các sinh viên MIT trong một khóa học mùa xuân năm ngoái được phát hiện lần đầu và sử dụng rộng rãi từ lâu trước khi các nhà vật lý học khẳng định sự tồn tại của nguyên tử và phân tử!

Ví dụ về sóng minh họa cho ba ý quan trọng. Thứ nhất, độc lập với bản chất vật lý không có nghĩa là ta không cần đến bản chất vật lý, mà là hầu hết chi tiết về nó không quan trọng. Rõ ràng bạn không thể có sóng âm trong chất khí nếu không có chất khí, nhưng đó là chất khí nào cũng được. Tương tự, rõ ràng bạn không thể có sự tính toán mà không có vật chất, nhưng đó là vật chất nào cũng được miễn là nó có thể sắp xếp thành các cổng NAND, các nơ-ron liên kết với nhau hay một thành phần cơ bản nào đó xây dựng nên phép tính toán phổ quát. Thứ hai, hiện tượng này độc lập với bản chất vật lý và tự nó có sự sống riêng, không phụ thuộc vào bản chất vật lý. Sóng có thể đi qua mặt hồ, cho dù không hề có phân tử nước nào di chuyển—hầu hết chúng chỉ nảy lên nảy xuống, như những người hâm mộ đứng lên ngồi xuống “tạo sóng” trên khán đài sân vận động. Thứ ba, ta thường chỉ quan tâm đến khía cạnh không liên quan đến bản chất

vật lý: một người lướt sóng thường quan tâm tới vị trí và độ cao của con sóng hơn là cấu tạo phân tử chi tiết của con sóng đó. Ta đã thấy điều này đúng với thông tin và nó cũng đúng với tính toán: nếu hai lập trình viên cùng nhau vạch mã tìm lỗi, có lẽ họ sẽ chẳng cần bàn đến các transistor.

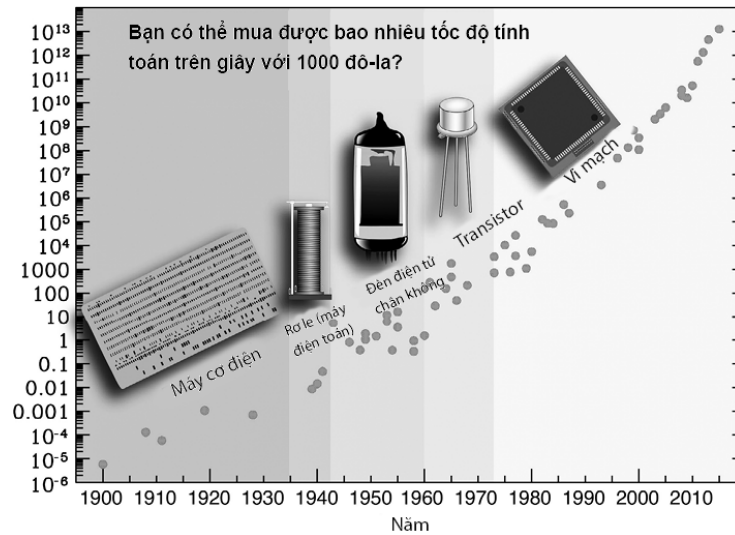
Giờ ta đã đến với câu trả lời cho câu hỏi mở đầu về việc làm thế nào những vật thể hữu hình có thể hình thành nên một thứ có cảm giác phi vật thể, trừu tượng và hư vô như trí tuệ: ta cảm giác nó phi vật chất như vậy là vì nó độc lập với bản chất vật lý, tự nó có sự sống riêng không phụ thuộc hay phản ánh những chi tiết vật lý. Nói ngắn gọn, sự tính toán là một khuôn mẫu sắp xếp phần tử trong không gian-thời gian, và không phải các phần tử đó quan trọng mà chính cái khuôn để sắp xếp các phần tử đó mới thật sự quan trọng! Vật chất không ảnh hưởng gì.

Nói cách khác, phần cứng là vật chất còn phần mềm là khuôn mẫu. Sự độc lập với bản chất vật lý của tính toán ngụ ý rằng AI là điều hoàn toàn khả thi: trí tuệ không cần đến máu thịt hay các phân tử cacbon mới tồn tại được.

Bởi có sự độc lập với bản chất vật lý ấy, các kỹ sư hăng hái mới có thể liên tục thay thế công nghệ bên trong máy tính bằng những công nghệ tốt hơn rất nhiều lần, mà không cần thay đổi phần mềm. Kết quả với cải thiện bộ nhớ cũng tuyệt vời hết như vậy. Như minh họa trong hình 2.8, phần cứng tính toán cũng rẻ đi một nửa sau mỗi hai năm, và xu hướng này đã kéo dài suốt hơn một thế kỷ qua, giảm giá thành cho máy tính đến một triệu triệu triệu (10^{18}) lần kể từ thời bà nội bà ngoại tôi ra đời. Nếu mọi thứ có giá rẻ hơn đến một triệu triệu triệu lần thì với một phần trăm của một xu Mỹ bạn có thể mua được tất cả hàng hóa dịch vụ làm ra trên Trái Đất trong năm nay. Dĩ nhiên, sự giảm giá ngoạn mục này là lý do chính khiến tính toán phổ biến khắp nơi như hiện nay; nó đã lan tỏa từ những cơ sở tính toán có kích thước cả tòa nhà ngày xưa đến từng ngôi nhà, chiếc xe và túi áo túi quần—thậm chí xuất hiện ở những nơi không ngờ như trong đế giày.

Tại sao công nghệ cứ tiếp tục nhân đôi năng lực theo chu kỳ đều đặn, thể hiện cái mà các nhà toán học gọi là tăng trưởng theo cấp số

nhân?



Hình 2.8: Từ năm 1900, cứ mỗi hai năm việc tính toán lại giảm giá thành một nửa. Bảng này cho thấy năng lực tính toán đo bằng số lượng phép toán dấu phẩy động trên giây (FLOPS) có thể mua được với 1000 đô-la Mỹ.³ Phép toán dấu phẩy động là tính toán tương ứng với 105 các phép toán logic cơ bản như phép toán thao tác bit hay đánh giá NAND.

Tại sao chuyện này lại xảy ra không chỉ về mặt thu nhỏ transistor (tạo ra xu hướng định luật Moore) mà còn rộng hơn là đối với tính toán nói chung (hình 2.8), bộ nhớ (hình 2.4) và rất nhiều công nghệ khác từ lập bản đồ gen đến chiếu chụp thần kinh? Ray Kurzweil gọi hiện tượng nhân đôi liên tiếp này là “định luật lợi nhuận tăng tốc”.

Tất cả ví dụ về việc nhân đôi không ngừng mà tôi biết trong tự nhiên đều có nguyên do cơ bản như nhau, và hiện tượng công nghệ này cũng không phải ngoại lệ: mỗi bước tạo ra bước tiếp theo. Ví dụ, bản thân bạn cũng trải qua tăng trưởng cấp số nhân ngay sau khi phôi thai bạn hình thành: mỗi tế bào của bạn chia đôi và sinh ra hai tế bào mới gần như mỗi ngày, khiến tổng số lượng tế bào của bạn tăng lên từng ngày từ 1 lên 2, 4, 8, 16 và cứ như thế. Theo lý thuyết khoa học phổ biến nhất về nguồn gốc vũ trụ, gọi là lạm phát vũ trụ, Vũ trụ sơ sinh của chúng ta đã từng tăng trưởng theo cấp số nhân cũng như bạn vậy, liên tục lặp lại việc nhân đôi kích thước theo chu kỳ đều đặn cho tới khi một hạt nhỏ hơn và nhẹ hơn cả nguyên tử lại phình ra

lớn hơn tất cả những thiên hà ta từng thấy được bằng kính viễn vọng. Một lần nữa, nguyên nhân là một quá trình trong đó mỗi bước nhân đôi dẫn đến bước tiếp theo. Đó cũng là cách công nghệ phát triển: một khi công nghệ trở nên mạnh gấp đôi, ta thường hay dùng công nghệ đó để lại thiết kế và xây dựng công nghệ mới mạnh gấp đôi, chêm ngòi cho một quá trình nhân đôi khả năng lặp đi lặp lại theo tinh thần định luật Moore.

Một điều xảy ra cũng thường xuyên như việc nhân đôi năng lực công nghệ đó là những lời khẳng định rằng quá trình nhân đôi đang dừng lại. Phải, định luật Moore tất nhiên sẽ có hồi kết, tức là sẽ có một giới hạn vật lý cho khả năng ta thu nhỏ transistor. Nhưng một số người lại nhầm rằng định luật Moore đồng nghĩa với sự nhân đôi liên tiếp năng lực công nghệ. Trái lại, Ray Kurzweil chỉ ra rằng định luật Moore không liên quan đến loại hình công nghệ thứ nhất mà là loại hình thứ năm trong quá trình đưa đến tăng trưởng cấp số nhân trong tính toán, như minh họa trong hình 2.8: bất cứ khi nào một công nghệ ngừng cải tiến, ta lại thay thế nó bằng một cái tốt hơn. Khi ta không thể thu nhỏ kích cỡ đèn điện tử chân không nữa, ta thay thế nó bằng transistor và sau đó là vi mạch, nơi các electron di chuyển trong không gian hai chiều. Khi công nghệ này đạt đến giới hạn, còn có rất nhiều biện pháp thay thế mà ta có thể thử—ví dụ như sử dụng mạch ba chiều và dùng thứ gì đó không phải là electron để thực hiện công việc.

Không ai biết chắc vật liệu tính toán mang tính đột phá tiếp theo sẽ là gì, nhưng chúng ta đều biết rằng còn lâu ta mới đến được giới hạn của các định luật vật lý. Cộng sự của tôi ở MIT, Seth Lloyd, đã tìm hiểu về giới hạn cơ bản này và các bạn sẽ khám phá chi tiết hơn trong chương 6 rằng giới hạn này vượt xa gấp 10^{33} lần so với công nghệ tiên tiến nhất hiện nay theo khả năng tính toán của một khối vật chất. Vậy nên, cho dù ta có tiếp tục nhân đôi năng lực của máy tính hai năm một lần thì cũng sẽ phải mất hơn hai thế kỷ nữa ta mới chạm đến ranh giới cuối cùng đó.

Mặc dù tất cả máy tính phổ quát đều có khả năng thực hiện các tính toán như nhau, nhưng một số máy tính có hiệu suất cao hơn. Chẳng hạn, một máy tính cần đến hàng triệu phép nhân không yêu

cầu phải có hàng triệu mô-đun phép nhân riêng biệt làm từ các transistor riêng biệt như trong hình 2.6: nó chỉ cần một mô-đun như vậy, vì nó có thể dùng mô-đun đó nhiều lần liên tiếp với đầu vào phù hợp. Để có được hiệu suất như vậy, hầu hết các máy tính hiện đại sử dụng một mô hình chia nhỏ các phép tính toán thành nhiều bước thời gian; trong các bước này, thông tin được chuyển qua chuyển lại giữa các mô-đun bộ nhớ và mô-đun tính toán. Kiến trúc tính toán này được phát triển từ những năm 1935 đến 1945 bởi các nhà tiên phong về máy tính bao gồm Alan Turing, Konrad Zuse, Presper Eckert, John Mauchly và John von Neumann. Cụ thể, bộ nhớ máy tính này lưu trữ cả dữ liệu và phần mềm (một chương trình, tức là một danh sách các chỉ lệnh có nhiệm vụ xử lý dữ liệu). Ở mỗi bước thời gian, một bộ xử lý trung tâm (CPU) thi hành chỉ lệnh tiếp theo trong chương trình; lệnh đó áp dụng một hàm số đơn giản lên một phần nào đó của dữ liệu. Bộ phận máy tính theo dõi công việc tiếp theo chỉ là một phần khác trong bộ nhớ, gọi là bộ đếm chương trình, lưu trữ số thứ tự của dòng hiện tại trong chương trình. Để đi đến chỉ dẫn tiếp theo, chỉ cần thêm một vào bộ đếm chương trình. Để nhảy đến một dòng khác trong chương trình, chỉ cần sao chép số thứ tự dòng đó vào bộ đếm chương trình—đó là cách những câu lệnh ‘if’ (nếu) và vòng lặp được thực hiện.

Máy tính ngày nay thường được tăng tốc nhờ xử lý song song, điều hơi ngược so với việc sử dụng lại một mô-đun nhưng theo cách rất thông minh: nếu có thể chia nhỏ phép toán thành nhiều phần có thể làm song song (nếu đầu vào của phần này không phụ thuộc đầu ra của phần kia), thì các phần đó có thể được tính toán cùng lúc bằng các bộ phận phần cứng khác nhau.

Dạng máy tính song song tối ưu là máy tính lượng tử. Nhà tiên phong trong tính toán lượng tử David Deutsch từng đưa ra luận điểm gây tranh cãi rằng “máy tính lượng tử chia sẻ thông tin với một số lượng khổng lồ các phiên bản của chính nó khắp đa vũ trụ”, và có thể tính toán ở Vũ trụ của chúng ta nhanh hơn nhờ sự hỗ trợ các phiên bản khác theo cách nhìn nào đó.⁴ Ta vẫn chưa biết có thể chế tạo được một máy tính lượng tử giá thành phải chăng trong những thập kỷ tới không, vì điều này phụ thuộc vào cả việc vật lý lượng tử có vận

hành như cách chúng ta nghĩ không và liệu khả năng của chúng ta có thể vượt qua những thách thức kỹ thuật to lớn, nhưng các doanh nghiệp và chính phủ trên khắp thế giới đang đánh cược hàng chục triệu đô-la Mỹ mỗi năm vào khả năng này. Tuy máy tính lượng tử không thể tăng tốc việc thực hiện những phép toán thông thường, nhưng người ta đã phát triển những thuật toán thông minh để có thể cải thiện đáng kể tốc độ một số dạng phép toán cụ thể, như giải mã các hệ thống mật mã và đào tạo các mạng nơ-ron. Máy tính lượng tử còn có thể mô phỏng hiệu quả về hành vi của các hệ thống cơ học lượng tử, bao gồm nguyên tử, phân tử và các vật chất mới, thay thế việc đo lường trong phòng thí nghiệm hóa học y như cách mà mô phỏng trên máy tính truyền thống đã thay thế việc đo lường trong đường hầm gió.

Học tập là gì?

Mặc dù một chiếc máy tính bỏ túi có thể đánh bại tôi trong một bài kiểm tra số học, nó sẽ không thể nào cải thiện tốc độ hay độ chính xác cho dù có luyện tập đến bao nhiêu. Nó không học tập: ví dụ, mỗi lần tôi ấn nút căn bậc hai, nó đều tính toán đúng hàm số căn bậc hai đó mà không thay đổi. Tương tự, chương trình máy tính đầu tiên đã đánh bại tôi trong môn cờ vua không rút ra bài học từ những lỗi sai của nó, mà chỉ thực hiện một hàm số mà lập trình viên khôn ngoan của nó đã thiết kế để tính toán nước đi tiếp theo thật hay. Ngược lại, khi Magnus Carlsen thua trận cờ vua đầu tiên hồi năm tuổi, anh đã bắt đầu một quá trình học tập để đến với đỉnh cao Vô địch Cờ vua Thế giới 18 năm sau đó.

Có thể nói, khả năng học tập là khía cạnh kỳ thú nhất của trí tuệ nói chung. Chúng ta đã thấy khả năng ghi nhớ và tính toán của một khối vật chất tưởng như vô tri, nhưng nó học tập bằng cách nào? Ta đã thấy việc trả lời một câu hỏi khó tương ứng với hoạt động tính toán một hàm số, và nếu vật chất được sắp xếp thích hợp thì nó có thể tính toán bất kỳ hàm số tính toán được nào. Khi con người chúng ta mới tạo ra máy tính bỏ túi và chương trình đánh cờ vua, chúng ta đã thực hiện việc sắp xếp đó. Để vật chất có thể học tập thì nó phải tự biết sắp xếp lại bản thân để ngày càng giỏi hơn trong việc tính

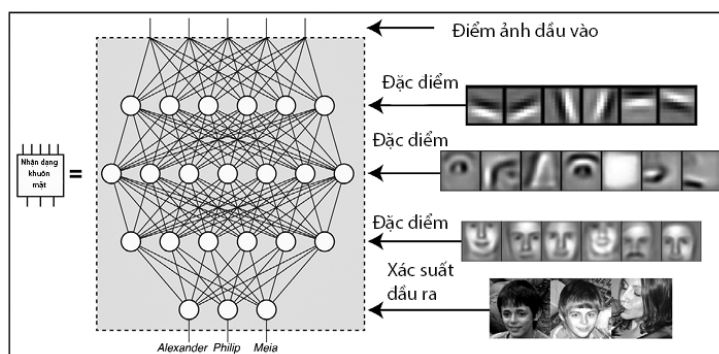
toán cùng một hàm số đã định—chỉ bằng cách tuân thủ các định luật vật lý.

Để giải mã bí ẩn của quá trình học tập, đầu tiên ta hãy xem xét cách một hệ thống vật lý rất đơn giản học các chữ số trong số π và các số khác. Ở trên ta đã thấy một bề mặt có nhiều thung lũng (xem hình 2.3) có thể dùng làm công cụ ghi nhớ: ví dụ, nếu đáy của một trong các thung lũng có vị trí $x = \pi \approx 3.14159$ và không có thung lũng nào khác gần đó, thì bạn có thể đặt bóng ở vị trí $x = 3$ và nhìn hệ thống tự tính ra những chữ số còn thiếu sau dấu thập phân khi quả bóng lăn dần xuống đáy. Giờ, giả sử bề mặt được làm từ đất sét mềm và ban đầu hoàn toàn phẳng phiu như một phiến đá trơn. Nếu có những người đam mê môn toán liên tục đặt quả bóng vào vị trí các con số yêu thích của họ, thì trọng lực sẽ dần dần tạo ra thung lũng ở các vị trí này, sau đó bề mặt đất sét có thể được sử dụng để nhớ lại những ký ức đã được lưu trữ. Nói cách khác, bề mặt đất sét đã học tập cách tính những chữ số của các số như π .

Các hệ thống vật lý khác như não bộ có thể học tập hiệu quả hơn nhiều dựa trên chính phương thức này. John Hopfield cho thấy mạng lưới nơ-ron liên kết với nhau nêu trên có thể học tập theo cách tương tự: nếu bạn liên tục đặt nó vào những trạng thái nhất định, nó sẽ dần dần học tập những trạng thái này và trở lại trạng thái đó từ bất kỳ trạng thái nào lân cận. Nếu bạn đã gặp các thành viên trong gia đình mình nhiều lần, thì ký ức về mặt mũi họ ra sao có thể được gọi lại bằng bất cứ thứ gì liên quan đến họ.

Mạng nơ-ron giờ đây đã biến đổi cả trí tuệ sinh học và nhân tạo, và gần đây bắt đầu thống trị một ngành con của AI gọi là học máy (nghiên cứu các thuật toán tự cải tiến qua kinh nghiệm thu được). Trước khi đi sâu hơn vào phương thức học tập của các mạng lưới như vậy, ta nên hiểu cách thức mà chúng tính toán. Một mạng nơ-ron đơn giản là một nhóm các nơ-ron (tế bào thần kinh) liên kết có khả năng ảnh hưởng đến hành vi của nhau. Não bạn chứa số lượng nơ-ron nhiều như số sao trong thiên hà của chúng ta: khoảng một trăm tỷ. Trung bình mỗi nơ-ron này liên kết với khoảng một ngàn nơ-ron khác qua những cầu nối gọi là xi-nap, và chính độ mạnh yếu của khoảng một trăm ngàn tỷ liên kết xi-nap này đã giúp mã hóa hầu hết thông tin trong não bộ.

Ta có thể vẽ sơ đồ mạng nơ-ron dưới dạng một tập hợp các chấm tượng trưng cho nơ-ron được liên kết bởi những đường thẳng tượng trưng cho xi-nap (xem hình 2.9). Trên thực tế, nơ-ron là những thiết bị điện hóa rất phức tạp không hề giống như sơ đồ minh họa này: chúng bao gồm nhiều bộ phận với nhiều cái tên như sợi trục và sợi nhánh, có nhiều loại nơ-ron hoạt động theo rất nhiều cách khác nhau, và chi tiết chính xác về việc xung điện từ một nơ-ron tác động đến các nơ-ron khác khi nào và thế nào vẫn là chủ đề đang được nghiên cứu. Tuy nhiên, các nhà nghiên cứu AI đã cho thấy mạng nơ-ron vẫn có thể đạt được trình độ con người ở nhiều nhiệm vụ hết sức phức tạp, ngay cả khi thay thế các nơ-ron sinh học thực tế phức tạp và chi tiết bằng những nơ-ron mô phỏng cực kỳ đơn giản, giống hệt nhau và tuân theo những quy tắc đơn giản. Mô hình mạng nơ-ron nhân tạo phổ biến nhất hiện nay gán cho mỗi trạng thái nơ-ron một con số duy nhất và độ mạnh của mỗi xi-nap một con số duy nhất. Trong mô hình này, mỗi nơ-ron cập nhật trạng thái của mình ở từng bước thời gian đều đặn chỉ bằng cách tính trung bình các đầu vào từ tất cả các nơ-ron kết nối với nó, tính trọng số dựa trên độ mạnh của xi-nap, có thể cộng thêm một hằng số, rồi ứng dụng hàm số kích hoạt cho kết quả để tính toán trạng thái tiếp theo của nó.²¹ Cách dễ nhất để dùng mạng nơ-ron như một hàm là khiến nó truyền thẳng, tức chỉ để thông tin truyền đi một chiều, như trong hình 2.9, nhập đầu vào cho hàm số ở lớp nơ-ron trên cùng và lấy đầu ra từ lớp nơ-ron dưới cùng.



Hình 2.9: Một mạng nơ-ron cũng có thể tính toán các hàm số như mạng lưới các cổng NAND. Ví dụ, các mạng nơ-ron nhân tạo được

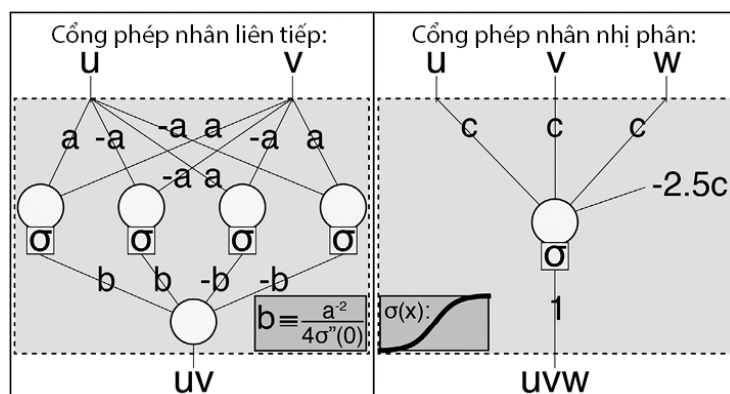
đào tạo để nhập đầu vào là các số tượng trưng cho độ sáng của các điểm ảnh khác nhau và đầu ra là các số tượng trưng cho xác suất hình ảnh đó mô tả người nào trong số nhiều người khác nhau. Ở đây mỗi nơ-ron nhân tạo (hình tròn) tính toán tổng có trọng số của các con số gửi đến nó qua các kết nối (đường thẳng) từ tầng trên, áp dụng một hàm đơn giản và chuyển kết quả xuống tầng dưới, cứ xuống mỗi tầng lại tính toán những đặc điểm tổng quát hơn. Các mạng lưới nhận dạng khuôn mặt thường chứa hàng trăm ngàn nơ-ron; hình minh họa này chỉ cho thấy một số nhỏ để các bạn dễ theo dõi.

Thành công của các mạng nơ-ron nhân tạo đơn giản này một lần nữa lại là ví dụ của sự độc lập với bản chất vật lý: mạng nơ-ron có khả năng tính toán mạnh có vẻ không phụ thuộc vào những tiểu tiết trong cấu tạo của nó. Trên thực tế, George Cybenko, Kurt Hornik, Maxwell Stinchcombe và Halbert White đã chứng minh một điều đáng chú ý vào năm 1989: những mạng nơ-ron đơn giản như vậy mang tính phổ quát, tức là chúng có thể tính toán đủ chính xác bất kỳ hàm số nào, chỉ bằng cách điều chỉnh các thông số qua độ mạnh của xi-nap. Nói cách khác, có vẻ quá trình tiến hóa không khiến nơ-ron sinh học của chúng ta trở nên phức tạp vì cần phức tạp, mà vì như vậy sẽ hiệu quả hơn—và bởi vì trái với các kỹ sư loài người, quá trình tiến hóa không tưởng thưởng cho những thiết kế đơn giản và dễ hiểu.

Khi mới biết điều này, tôi cảm thấy thật bí ẩn khi một thứ đơn giản như thế có thể tính toán một thứ phức tạp như vậy. Ví dụ, làm sao bạn có thể tính toán một thứ đơn giản như phép nhân, khi bạn chỉ có thể tính tổng có trọng số và ứng dụng một hàm số ấn định duy nhất? Trong trường hợp bạn muốn tìm hiểu đôi chút về cách vận hành này, hình 2.10 cho thấy chỉ cần năm nơ-ron là có thể nhân hai số ngẫu nhiên với nhau và một nơ-ron duy nhất có thể nhân ba bit với nhau.

Tuy có thể chứng minh rằng về lý thuyết ta có thể tính toán bất kỳ thứ gì với một mạng nơ-ron đủ lớn, nhưng bằng chứng đó không thể hiện khả năng bạn có thể làm việc này trên thực tế, với một mạng nơ-ron có kích cỡ hợp lý. Càng suy ngẫm, tôi lại càng thấy khó hiểu vì sao mạng nơ-ron lại hoạt động hiệu quả đến như vậy.

Giả sử, ta muốn phân loại các hình ảnh đen trắng với tổng số điểm ảnh là một triệu thành hai loại là chó hoặc mèo. Nếu mỗi điểm ảnh có thể nhận một trong 256 giá trị, thì có đến $256^{1000000}$ khả năng hình ảnh khả thi; ở mỗi hình ảnh, ta muốn tính toán xác suất đó là hình ảnh một con mèo. Đây là một hàm số bất kỳ nhận đầu vào là một hình ảnh và đầu ra là một xác suất, được định nghĩa bởi một danh sách gồm $256^{1000000}$ khả năng. Đây là con số lớn hơn rất nhiều so với số lượng nguyên tử trong Vũ trụ chúng ta (khoảng 10^{78}). Vậy mà mạng nơ-ron chỉ với hàng ngàn hay triệu thông số lại có thể thực hiện nhiệm vụ phân loại như vậy khá tốt. Làm thế nào mạng nơ-ron thành công lại có thể “rẻ” như vậy, tức là chỉ cần đến ít thông số như vậy? Suy cho cùng, bạn vẫn có thể chứng minh một mạng nơ-ron đủ nhỏ để nằm vừa bên trong Vũ trụ sẽ thất bại thảm hại trong việc thực hiện gần như tất cả các hàm số và chỉ thành công ở một phần nhỏ xíu đến mức cười nhạo những nhiệm vụ tính toán mà bạn định giao cho nó.



Hình 2.10: Thay vì bằng NAND như trong hình 2.7, đây là cách vật chất thực hiện phép nhân bằng nơ-ron. Bạn không cần đi vào chi tiết mà chỉ cần biết là nơ-ron (nhân tạo hay sinh học) có thể làm toán và phép nhân yêu cầu lượng nơ-ron ít hơn số cổng NAND rất nhiều. Thông tin thêm dành cho người cuồng môn toán: Các hình tròn thực hiện tính tổng, hình vuông ứng dụng hàm σ , các đường thẳng làm phép nhân với hằng số ghi trên đường thẳng đó. Đầu vào là các số thực (trái) và các bit (phải). Phép nhân trở nên chính xác tùy ý khi $a \rightarrow 0$ (trái) và $c \rightarrow \infty$ (phải). Mạng bên trái có hiệu quả với bất kỳ hàm $\sigma(x)$ nào có đoạn đầu là đường cong (với đạo hàm bậc hai $\sigma''(0) \neq 0$),

có thể chứng minh bằng cách khai triển chuỗi Taylor cho hàm $\sigma(x)$. Mạng bên phải yêu cầu hàm $\sigma(x)$ tiến đến 0 và 1 khi giá trị x lần lượt cực nhỏ và cực lớn, có thể thấy $uvw = 1$ chỉ khi $u + v + w = 3$. (Các ví dụ này lấy từ một bài nghiên cứu tôi viết cùng các sinh viên của tôi, Henry Lin và David Rolnick, “Why Does Deep and Cheap Learning Work So Well?” (Vì sao học sâu và rẻ lại hiệu quả đến vậy?) có thể tìm trên trang <http://arxiv.org/abs/1608.08225>.) Bằng cách kết hợp nhiều phép nhân (như trên) và phép cộng, bạn có thể tính toán bất kỳ đa thức nào, và như nhiều người biết, đa thức có thể xấp xỉ bất kỳ hàm số trơn nào.

Tôi đã có nhiều niềm vui khi cùng đầu đầu khám phá vấn đề này và các bí ẩn liên quan với Henry Lin, sinh viên của tôi. Một trong những điều tôi thấy biết ơn nhất trong đời là cơ hội được cộng tác với những sinh viên tuyệt vời, và Henry là một trong số đó. Khi cậu lần đầu bước vào văn phòng tôi để hỏi xem tôi có hứng thú làm việc với cậu không, tôi đã thầm nghĩ đúng ra tôi phải hỏi cậu có hứng thú làm việc với tôi không: cậu nhóc mắt sáng, thân thiện và khiêm tốn quê ở Shreveport, Louisiana này đã viết được tám bài nghiên cứu khoa học, giành danh hiệu Forbes 30-Under-30²², và là diễn giả của một buổi diễn thuyết TED được hơn một triệu lượt xem—ở tuổi 20! Một năm sau, chúng tôi cùng nhau viết một bài nghiên cứu với kết luận bất ngờ: câu hỏi về hiệu quả của mạng nơ-ron không thể chỉ trả lời bằng toán học, vì một phần câu trả lời nằm ở vật lý học. Chúng tôi thấy rằng loại hàm số mà các định luật vật lý đưa ra cho ta và khiến ta quan tâm tới tính toán là một loại cực kỳ nhỏ vì, do những nguyên nhân chúng ta chưa hiểu rõ, các định luật vật lý đơn giản đến khó tin. Hơn nữa, phần rất nhỏ các hàm số mà mạng nơ-ron có thể tính toán được lại rất tương đồng với phần nhỏ các hàm số mà vật lý học khiến ta quan tâm tới! Chúng tôi cũng mở rộng các công trình trước đó, cho thấy các mạng nơ-ron học sâu (gọi là “sâu” khi phải qua nhiều tầng) hiệu quả hơn nhiều những mạng nông đối với các hàm số cần quan tâm kia. Ví dụ, cùng với một sinh viên MIT tài ba khác là David Rolnick, chúng tôi đã cho thấy nhiệm vụ đơn giản là nhân n số cần những $2n$ nơ-ron trong mạng nông chỉ có một lớp, nhưng chỉ cần khoảng $4n$ nơ-ron trong mạng sâu. Điều này giúp giải thích không chỉ

vì sao mạng nơ-ron giờ lại thịnh hành trong giới nghiên cứu AI đến vậy, mà còn cả vì sao chúng ta lại tiến hóa mạng nơ-ron trong chính não bộ: nếu chúng ta tiến hóa bộ não để có thể dự đoán tương lai, thì cũng hợp lý khi ta thực hiện tiến hóa một cấu trúc tính toán giải những vấn đề tính toán quan trọng trong thế giới vật lý.

Giờ ta đã khám phá cách các mạng nơ-ron hoạt động và tính toán, hãy trở lại với câu hỏi rằng chúng học như thế nào. Cụ thể là, làm sao mạng nơ-ron có thể tính toán tốt hơn bằng cách cập nhật các xi-nap?

Trong cuốn sách tiền đề năm 1949, *The Organization of Behavior: A Neuropsychological Theory* (Tổ chức hành vi: Một lý thuyết thần kinh học), nhà tâm lý học người Canada Donald Hebb đã đưa ra luận điểm rằng nếu hai nơ-ron gần nhau thường xuyên kích hoạt (“đánh lửa”) cùng lúc, kết nối xi-nap của chúng sẽ mạnh lên khiến chúng học được cách kích hoạt lẫn nhau—ý tưởng được đúc kết trong câu nói nổi tiếng “Cùng đánh lửa, cùng kết nối” trong sách. Tuy chúng ta còn lâu mới hiểu chi tiết về cách thức học tập thực tế của não bộ, và nghiên cứu đã chỉ ra rằng trong nhiều trường hợp, các câu trả lời phức tạp hơn nhiều, nhưng nghiên cứu cũng đồng thời cho thấy chỉ bằng quy tắc học tập đơn giản này (gọi là Thuyết học tập Hebb) mạng nơ-ron cũng có thể học được nhiều điều thú vị. John Hopfield đã chỉ ra thuyết học tập Hebb cho phép mạng nơ-ron nhân tạo đơn giản hóa của ông lưu trữ được nhiều ký ức phức tạp chỉ bằng cách tiếp xúc nhiều lần với những ký ức đó. Sự tiếp xúc với thông tin cần học thường được gọi là “đào tạo” khi nói đến mạng nơ-ron nhân tạo (hay nói đến động vật hoặc con người được dạy cho các kỹ năng), mặc dù dùng từ “học tập”, “giáo dục” hay “trải nghiệm” đều thỏa đáng. Các mạng nơ-ron nhân tạo vận hành các hệ thống AI ngày nay thường thay thế thuyết học tập Hebb bằng các quy tắc học tập tinh vi hơn với những cái tên phức tạp như “truyền ngược” hay “phương pháp xuống dốc phân phối xác suất ngẫu nhiên”, nhưng ý tưởng cơ bản vẫn thế: có một quy tắc tắt định đơn giản và gắn bó với một định luật vật lý, theo đó các xi-nap được cập nhật qua thời gian. Như một phép màu, quy tắc đơn giản này có thể khiến mạng nơ-ron học những cách tính toán phức tạp đáng kể nếu việc đào tạo được thực hiện với lượng dữ liệu lớn. Tuy ta vẫn chưa biết chính xác các quy

tắc học tập mà bộ não con người sử dụng, nhưng dù câu trả lời là gì thì cũng không có biểu hiện nào cho thấy chúng vượt quá giới hạn của các định luật vật lý.

Cũng như hầu hết máy tính kỹ thuật số tăng cường hiệu quả bằng cách chia nhỏ công việc thành nhiều bước và sử dụng lại nhiều lần các mô-đun tính toán, các mạng nơ-ron sinh học và nhân tạo cũng vậy. Não bộ có những bộ phận mà các nhà khoa học máy tính gọi là mạng nơ-ron hồi quy thay vì truyền thẳng, trong đó thông tin có thể truyền đi theo nhiều hướng chứ không chỉ một chiều, nên đầu ra hiện tại có thể trở thành đầu vào cho bước tiếp theo. Mạng lưới các cổng logic trong bộ vi xử lý của máy tính cá nhân cũng hồi quy theo nghĩa này: nó liên tục tái sử dụng các thông tin cũ và để thông tin mới đi vào từ bàn phím, bàn di chuột, camera, v.v. tác động lên tính toán liên tục, từ đó xác định thông tin đầu ra ở màn hình, loa, máy in hay mạng không dây. Tương tự, mạng nơ-ron trong não bộ là hồi quy, để thông tin đầu vào từ mắt, tai và các giác quan khác tác động lên tính toán liên tục, từ đó xác định thông tin đầu ra cho các cơ.

Lịch sử sự học ít nhất cũng dài như lịch sử sự sống, vì mọi sinh vật tự sinh sản đều thực hiện quá trình thú vị là sao chép và xử lý thông tin—hành vi đã được học từ trước theo cách nào đó. Tuy nhiên, trong thời đại Sự sống 1.0 thì các sinh vật không học được gì ngay trong vòng đời của chúng: các quy tắc xử lý thông tin và phản ứng của chúng chỉ được xác định bởi gen di truyền, nên quá trình học duy nhất với chúng diễn ra rất chậm ở cấp độ loài, thông qua quá trình tiến hóa theo thuyết Darwin kéo dài nhiều thế hệ.

Khoảng nửa tỉ năm trước, một số dòng gen nhất định trên Trái Đất đã phát hiện ra cách khiến động vật phát triển mạng nơ-ron để có thể học được các hành vi từ những trải nghiệm trong cuộc sống. Sự sống 2.0 đã tới; khả năng học nhanh hơn nhiều giúp nó trở nên thông minh hơn đối thủ nên nó lan đi khắp địa cầu như đám cháy rừng. Như ta thấy ở chương 1, sự sống đã ngày càng giỏi hơn trong việc học tập, với một tốc độ ngày càng cao. Một loài giống như khi đã phát triển não bộ có khả năng thu thập kiến thức tài tình đến mức học được cách sử dụng công cụ, tạo ra lửa, nói được một ngôn ngữ và xây dựng một xã hội toàn cầu phức tạp. Bản thân xã hội này cũng có thể coi là một hệ thống biết ghi nhớ, tính toán và học tập; xã hội

này làm tất cả với một tốc độ ngày càng tăng, khi mà sáng chế này lại tạo điều kiện cho sáng chế tiếp theo: chữ viết, máy in ấn, khoa học hiện đại, máy tính, mạng internet, v.v.. Các sử gia tương lai sẽ viết gì thêm vào danh sách những sáng chế tạo ra tiền đề này? Tôi đoán họ sẽ viết về trí tuệ nhân tạo.

Như chúng ta đều biết, các cải tiến bùng nổ trong công nghệ bộ nhớ máy tính và năng lực tính toán (hình 2.4 và hình 2.8) đã trở thành tiến bộ đáng nể trong trí tuệ nhân tạo—nhưng phải mất một thời gian dài thì khái niệm học máy mới đạt đến độ chín. Khi máy tính Deep Blue của IBM thắng nhà vô địch cờ vua Garry Kasparov năm 1997, lợi thế lớn của nó nằm ở bộ nhớ và tính toán, không phải ở học tập. Trí tuệ máy tính của nó được tạo ra bởi một đội ngũ con người, và lý do chính khiến Deep Blue có thể thắng những người sáng tạo ra nó là vì nó tính toán nhanh hơn và do đó có khả năng phân tích những thế cờ tiềm năng hơn. Khi máy tính Watson của IBM hạ bệ nhà vô địch thế giới trong trò chơi đố vui Jeopardy!, nó cũng dựa trên khả năng học tập ít hơn dựa trên những kỹ năng đã lập trình riêng cho nó, hay bộ nhớ cũng như tốc độ ưu việt hơn. Ta cũng có thể nhận xét như vậy về những đột phá đầu tiên trong ngành chế tạo rô-bốt, từ vận động bằng chân đến xe tự lái và tên lửa tự hạ cánh.

Ngược lại, động lực đằng sau nhiều đột phá AI gần đây chính là học máy. Ví dụ, hãy xem hình 2.11. Để nói đó là hình ảnh gì thì dễ, nhưng để lập trình một hàm số có đầu vào không có gì ngoài màu sắc của tất cả các điểm ảnh trên hình và đầu ra là một dòng chú thích chính xác như “Một nhóm người trẻ chơi trò ném đĩa”, cũng đã khiến các nhà nghiên cứu AI trên toàn thế giới bó tay suốt nhiều thập kỷ. Khi đầu vào là một bộ màu sắc điểm ảnh khác thì hàm số trả lời “Một đàn voi đi qua đồng cỏ khô” lại là kết quả chính xác. Họ làm thế nào vậy? Theo kiểu Deep Blue, bằng cách lập trình những thuật toán thủ công để nhận dạng đĩa ném, các khuôn mặt và những thứ tương tự ư? Không, mà là bằng cách tạo ra một mạng nơ-ron tương đối đơn giản, không có chút hiểu biết gì về thế giới vật lý hay nội dung của nó, rồi để mạng nơ-ron này học bằng cách cho nó tiếp xúc với một lượng dữ liệu khổng lồ. Nhà tiên tri AI Jeff Hawkins đã viết năm 2004 rằng “không có máy tính nào có thể... nhìn giỏi bằng con chuột”, nhưng những ngày tháng đó giờ đã xa rồi.



Hình 2.11: “Một nhóm người trẻ chơi trò ném đĩa”—chú thích này được viết bởi một máy tính hoàn toàn không hiểu biết gì về người, trò chơi hay ném đĩa cả.

Cũng như ta không hiểu tường tận cách trẻ em học tập, ta không thể hiểu tường tận cách các mạng nơ-ron học tập và vì sao đôi khi chúng thất bại. Nhưng rõ ràng là giờ đây mạng nơ-ron học tập đã cực kỳ hữu dụng và đang chầm ngòi cho một cuộc chạy đua cho đầu tư vào học sâu. Học sâu giờ đã biến đổi nhiều phương diện của thị giác máy tính, từ chuyển chữ viết tay thành văn bản đến phân tích video thời gian thực dành cho xe tự lái. Tương tự, nó cũng đã đem lại một cuộc cách mạng ở khả năng của máy tính khi cho phép chuyển đổi ngôn ngữ nói thành văn bản và dịch ra ngôn ngữ khác ngay lập tức—thế nên bây giờ chúng ta mới có thể nói chuyện với những trợ lý cá nhân kỹ thuật số như Siri, Google Now và Cortana. Những câu đố CAPTCHA phiền phức mà chúng ta phải làm để thuyết phục một trang web rằng chúng ta là con người đang ngày càng khó vượt lên trên những gì công nghệ học máy có thể làm. Năm 2015, Google DeepMind phát hành một hệ thống AI sử dụng công nghệ học sâu có khả năng làm bá chủ trong hàng tá trò chơi điện tử như một đứa trẻ—mà không cần một chỉ dẫn nào—ngoại trừ việc nó nhanh chóng học được cách chơi giỏi hơn bất kỳ con người nào. Năm 2016, chính công ty đó cũng xây dựng nên AlphaGo, một hệ thống máy tính biết chơi cờ vây sử dụng học sâu để đánh giá sức mạnh của những thế cờ khác nhau và đánh bại nhà vô địch cờ vây mạnh nhất thế giới. Tiến bộ này đang tiếp nhiên liệu cho một quá

trình mở rộng, đem lại ngày càng nhiều nguồn vốn và nhân tài cho nghiên cứu AI, giúp tạo ra thêm nhiều tiến bộ.

Chúng ta đã dành chương này để khám phá bản chất và sự phát triển của trí tuệ cho tới ngày nay. Sẽ còn bao lâu nữa thì máy móc mới có thể vượt lên chúng ta trong mọi hoạt động nhận thức? Rõ ràng chúng ta không biết, và cũng nên để ngỏ khả năng cho câu trả lời “không bao giờ”. Tuy nhiên, thông điệp cơ bản trong chương này là chúng ta cũng cần cân nhắc khả năng điều đó sẽ xảy ra, thậm chí ngay trong cuộc đời chúng ta. Suy cho cùng, vật chất có thể được sắp xếp để khi tuân thủ các định luật vật lý, nó sẽ biết ghi nhớ, tính toán và học tập—và vật chất đó không nhất thiết cần có nguồn gốc sinh học. Các nhà nghiên cứu AI thường bị buộc tội là hứa hẹn thật nhiều rồi thất hứa cũng thật nhiều, nhưng công bằng mà nói, chính một số nhà phê bình đó cũng chưa đạt được nhiều công trạng là bao. Một số người luôn định nghĩa trí tuệ là thứ gì đó mà máy tính không thể làm được, hay thứ gì đó gây ấn tượng với nhân loại – điều này dường như đang liên tục đẩy xa mục tiêu của chúng ta. Máy móc giờ đã giỏi, hay phải nói là xuất sắc, trong các môn số học, cờ vua, chứng minh định lý toán học, chọn cổ phiếu, chú thích hình ảnh, lái xe, chơi trò chơi điện tử thùng, cờ vây, tổng hợp giọng nói, chuyển đổi giọng nói thành văn bản, dịch thuật và cả chẩn đoán ung thư, nhưng một số nhà phê bình vẫn sẽ khinh bỉ chế nhạo “Phải rồi—nhưng đó đâu phải trí tuệ thực sự!” Họ có thể tranh luận tiếp rằng trí tuệ thật sự chỉ bao gồm những đỉnh núi trong khung cảnh của Moravec (hình 2.2) mà tới giờ vẫn chưa ngập nước, cũng như một số người trong quá khứ từng đề cập về khả năng chú thích hình ảnh và cờ vây—trong khi nước vẫn tiếp tục dâng.

Nếu nước tiếp tục dâng trong một thời gian dài nữa, tác động của AI lên xã hội sẽ ngày càng lớn. Trước khi AI đạt tới trình độ con người trong mọi nhiệm vụ, nó sẽ đem lại cho chúng ta những cơ hội và thách thức hấp dẫn liên quan tới những vấn đề như lỗi kỹ thuật, luật pháp, vũ khí và công ăn việc làm. Những cơ hội và thách thức đó là gì và chúng ta có thể chuẩn bị tốt nhất bằng cách nào? Hãy tìm hiểu điều này trong chương tiếp theo.

TÓM TẮT:

Trí tuệ, định nghĩa là khả năng hoàn thành các mục tiêu phức tạp, không thể chỉ đo được bằng chỉ số IQ, mà phải bằng một phổ khả năng bao quát các mục tiêu.

Trí tuệ nhân tạo ngày nay thường hẹp, mỗi hệ thống chỉ có khả năng hoàn thành những mục tiêu rất cụ thể, trong khi trí tuệ con người rộng đến đáng kể.

Bộ nhớ, tính toán, học tập và trí tuệ có cảm giác trừu tượng, phi vật thể và hư vô vì chúng độc lập với bản chất vật lý: tức có khả năng tồn tại riêng biệt mà không phụ thuộc hay phản ánh những chi tiết của nguyên liệu nền chứa đựng chúng.

Bất kỳ khối vật chất nào cũng có thể làm nguyên liệu nền cho bộ nhớ miễn là nó có nhiều trạng thái ổn định khác nhau.

Bất kỳ vật chất nào cũng có thể là vật liệu thông minh, nguyên liệu nền cho tính toán, miễn là nó chứa những viên gạch phổ quát nhất định có khả năng kết hợp lại để thực hiện bất kỳ hàm số nào. Các cổng NAND và nơ-ron là hai ví dụ quan trọng cho những “nguyên tử tính toán” phổ quát như thế.

Một mạng nơ-ron là một nền vật lý mạnh mẽ cho quá trình học tập, vì chỉ bằng việc tuân thủ các định luật vật lý, nó có thể tự sắp xếp lại bản thân để ngày càng giỏi hơn trong việc thực hiện các tính toán mong muốn.

Vì các định luật vật lý rất đơn giản, nên con người chỉ quan tâm đến một phần rất nhỏ của tất cả các vấn đề tính toán có thể tưởng tượng ra được, và các mạng nơ-ron thường rất giỏi trong việc giải quyết chính phần nhỏ được quan tâm đó.

Một khi công nghệ mạnh gấp đôi, chúng ta thường có thể dùng nó để thiết kế và chế tạo công nghệ còn mạnh hơn nó gấp đôi, dẫn đến sự nhân đôi khả năng lặp đi lặp lại theo tinh thần định luật Moore. Giá thành của công nghệ thông tin trong vòng một thế kỷ qua cứ mỗi hai năm lại giảm đi một nửa, tạo điều kiện cho kỷ nguyên thông tin.

Nếu tiến bộ công nghệ AI tiếp tục, thì trước khi AI đạt đến trình độ con người ở tất cả các kỹ năng, nó sẽ đem lại cho ta những cơ hội và thách thức to lớn liên quan tới những vấn đề như lỗi

kỹ thuật, luật pháp, vũ khí và công ăn việc làm—chúng ta sẽ tìm hiểu những vấn đề này trong chương tiếp theo.

Chương 3

Tương Lai Gần: Đột phá, Lỗi, Luật, Vũ khí và Việc làm

Nếu chúng ta không đổi hướng sớm, chúng ta sẽ dừng chân ở nơi mình đang đến.

Irwin Corey

Con người cần phải ra sao trong thời đại ngày nay? Ví dụ, chúng ta coi trọng điều gì ở bản thân mình, và điều này khiến chúng ta khác biệt với các dạng sống khác và máy móc ra sao? Người khác coi trọng điểm gì ở chúng ta, đến mức sẵn lòng mời ta về làm việc? Cho dù ở thời điểm nào và câu trả lời là gì, chắc chắn sự phát triển công nghệ đã dần thay đổi các câu trả lời ấy.

Lấy tôi làm ví dụ. Là một nhà khoa học, tôi tự hào lập ra mục tiêu cá nhân, sử dụng óc sáng tạo và trực giác để xử lý một loạt vấn đề chưa được giải quyết và sử dụng ngôn ngữ để chia sẻ thứ tôi khám phá được. Cũng may cho tôi là xã hội sẵn sàng trả tiền để tôi làm công việc này. Còn nếu ở nhiều thế kỷ trước, giống như nhiều người khác, hẳn tôi đã phải gây dựng tên tuổi của mình bằng nghề làm nông hoặc thợ thủ công, nhưng sự phát triển của công nghệ từ đó đã khiến các ngành nghề này thu lại còn một phần nhỏ trong lực lượng lao động. Điều này có nghĩa là mọi người không còn khả năng xây dựng tên tuổi của họ qua nông nghiệp hoặc thủ công nghiệp nữa.

Về mặt cá nhân, tôi không phiền lòng khi thấy các cỗ máy ngày nay vượt trội hơn tôi ở các kỹ năng tay chân như đào bới và đan len, bởi đây không phải sở thích, nguồn thu nhập hay niềm tự hào của tôi. Thật vậy, mọi ảo tưởng của tôi về khả năng của mình trong lĩnh vực này đã bị nghiền nát hồi tôi tám tuổi, khi trường học buộc tôi phải tham gia lớp đan len mà tôi suýt nữa trượt và tôi chỉ hoàn thành bài tập nhờ sự giúp đỡ của một học sinh lớp năm thương hại tôi.

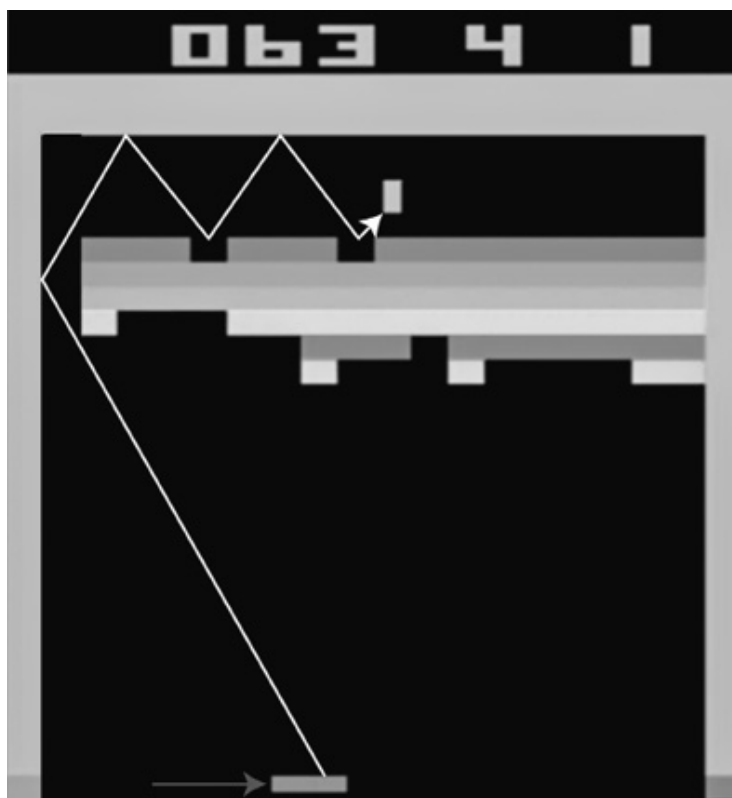
Nhưng khi công nghệ tiếp tục cải thiện, liệu sự trỗi dậy của AI có làm lu mờ cả những năng lực vẫn tạo ra cảm giác tự hào và giá trị hiện tại của tôi trên thị trường việc làm không? Stuart Russell kể với tôi rằng anh và nhiều người trong số các đồng nghiệp nghiên cứu AI

của anh gần đây đã trải nghiệm một khoảnh khắc “quý thần ơi!”, khi họ chứng kiến AI làm một việc mà họ không trông mong trong hàng nhiều năm tới. Trên tinh thần đó, xin hãy để tôi kể cho bạn nghe về vài khoảnh khắc “quý thần ơi!” của riêng tôi, và tại sao tôi coi chúng như điềm báo cho việc năng lực của con người sẽ sớm bị vượt qua.

Đột phá

Tác tử học sâu tăng cường

Vào năm 2014, tôi từng phải há hốc kinh ngạc nhiên khi xem video ghi lại cách hệ thống AI DeepMind²³ học cách chơi trò chơi máy tính. Cụ thể, AI này đang chơi trò Bắn tường (Breakout) (xem hình 3.1), một trò chơi Atari²⁴ cổ điển mà tôi rất yêu thích thuở thiếu niên. Mục tiêu của người chơi là điều khiển vợt sao cho bóng bật khỏi vợt lên bức tường gạch; mỗi khi bạn bắn trúng một viên gạch, nó biến mất và điểm số của bạn tăng lên.



Hình 3.1: Sau khi học chơi trò Bắn tường Atari từ đầu và ứng dụng khả năng học sâu tăng cường để tối đa hóa điểm số, AI DeepMind đã khám phá ra chiến lược tối ưu: nó chọn một lỗ ở ngoài cùng bên trái của bức tường gạch và để bóng tự nảy xung quanh phía sau tường, khiến điểm tích lũy rất nhanh. Tôi đã vẽ mũi tên thể hiện quỹ đạo trước đó của quả bóng và cái vợt.

Ngày trước, tôi từng tự mình viết một số trò chơi máy tính và biết rõ rằng viết một chương trình có thể chơi Bắn tường không khó – nhưng nhóm DeepMind đã không làm vậy. Thay vào đó, họ tạo ra một AI nguyên sơ không biết gì về trò chơi này – hoặc về bất kỳ trò chơi nào khác, thậm chí không biết đến các khái niệm như trò chơi, vợt, viên gạch hay quả bóng. AI của họ chỉ biết một danh sách dài các con số được đưa vào đều đặn gồm điểm số hiện tại và một dãy dài các con số mà chúng ta (chứ không phải AI) sẽ nhận ra là thông số kỹ thuật về màu sắc của các phần khác nhau thuộc màn hình. AI chỉ được yêu cầu đơn giản là tối đa hóa điểm số bằng cách đều đặn xuất ra con số mà chúng ta (chứ không phải AI) sẽ nhận ra là mã cho các phím nên bấm.

Ban đầu, AI chơi rất tệ: nó lúng túng đưa đẩy cái vợt trái phải một cách ngẫu nhiên và gần như luôn để rơi bóng. Sau một thời gian, có vẻ nó nhận thấy di chuyển vợt về phía quả bóng là ý tưởng hay, mặc dù phần lớn thời gian nó vẫn đỡ hụt bóng. Nhưng nó tiếp tục cải thiện qua thực hành và sớm trở nên giỏi trò chơi này hơn cả tôi; nó không bao giờ đỡ trượt bóng bất kể tốc độ nhanh đến mức nào. Và rồi tôi sửng sốt: nó đã tìm ra một chiến lược tối đa hóa điểm số tuyệt vời khi luôn nhắm đến góc trên bên trái để chọn một lỗ xuyên qua tường và làm cho quả bóng mắc kẹt nảy bật qua bật lại liên tục trong khoảng giữa bức tường và rào chắn sau nó. Đây có vẻ là một hành động thực sự thông minh. Quả thực, Demis Hassabis sau đó nói với tôi rằng nhiều lập trình viên trong nhóm DeepMind còn không biết mẹo này cho đến khi họ học được nó từ hệ thống AI họ xây dựng. Tôi khuyên bạn nên tận mắt xem video này tại đường link tôi đã cung cấp.¹

Một yếu tố giống con người đã phát sinh ở đây khiến tôi bất an: trước mắt tôi là một AI có mục tiêu, đang học tập để ngày càng giỏi

hơn trong việc đạt được mục tiêu đó, thậm chí vượt trội hơn cả những người sáng tạo ra nó. Trong chương trước, chúng ta đã định nghĩa trí tuệ là khả năng hoàn thành các mục tiêu phức tạp. Theo ý này, AI của DeepMind đã trở nên thông minh hơn ngay trước mắt tôi (mặc dù chỉ ở góc độ rất hẹp là chơi trò chơi cụ thể này). Trong chương đầu tiên, chúng ta đã gặp thứ các nhà khoa học máy tính gọi là tác tử thông minh tức các tổ chức thu thập thông tin về môi trường xung quanh qua cảm biến, sau đó xử lý thông tin này để quyết định cách chúng phản hồi. Mặc dù AI chơi game của DeepMind sống trong một thế giới ảo cực kỳ đơn giản chỉ gồm gạch, vợt và bóng, tôi không thể phủ nhận rằng nó là một tác tử thông minh.

DeepMind nhanh chóng công bố phương pháp và chia sẻ thuật toán của họ, rằng họ đã sử dụng ý tưởng rất đơn giản nhưng mạnh mẽ là học sâu tăng cường.² Học tăng cường cơ bản là một kỹ thuật học máy cổ điển bắt nguồn từ môn tâm lý học hành vi theo đó, việc nhận được phần thưởng tích cực sẽ khiến bạn làm một điều gì nhiều hơn và ngược lại. Cũng như việc một chú chó học diễn trò để tăng cơ hội được chủ khen ngợi hoặc thưởng bánh, AI của DeepMind đã học cách di chuyển vợt để bắt bóng bởi điều này sẽ tăng cơ hội ghi nhiều điểm nhanh hơn. DeepMind đã kết hợp ý tưởng này với học sâu: như trong chương trước, họ đã đào tạo một mạng lưới nơ-ron sâu để dự đoán số điểm trung bình sẽ đạt được từ việc nhấn từng phím cho phép trên bàn phím, để sau đó AI sẽ chọn phím bấm được mạng nơ-ron đánh giá là hứa hẹn nhất đối với trạng thái hiện tại của trò chơi.

Khi liệt kê các đặc điểm góp phần tạo nên cảm giác tự hào về giá trị riêng của bản thân, tôi cũng bao gồm cả khả năng xử lý hàng loạt vấn đề chưa được giải quyết. Ngược lại, việc chỉ biết chơi trò Bắn tường yêu cầu trí thông minh cực kỳ hạn hẹp. Nhưng với tôi, ý nghĩa thực sự trong đột phá của DeepMind nằm ở điểm rằng: học sâu tăng cường là một kỹ thuật cực kỳ tổng quát. Và họ đã cho chính AI này luyện 49 trò chơi Atari khác nhau, để rồi vượt qua người chơi ở 29 trò trong số đó, từ Pong (Bóng bàn) đến Boxing (Đấm bốc), Video Pinball (Bắn bi điện tử) và Space Invaders (Kẻ xâm lăng Ngoài vũ trụ).

Không lâu sau, ý tưởng được áp dụng để các AI này thể hiện năng lực ở những trò chơi hiện đại hơn trong thế giới ba chiều thay vì hai chiều. Chẳng mấy chốc, đối thủ của DeepMind là OpenAI ở San Francisco đã phát hành nền tảng Universe (Vũ Trụ), nơi AI của DeepMind và các tác tử thông minh khác có thể tương tác với toàn bộ một máy tính như thể đây là trò chơi: chúng có thể nhấn chuột vào bất cứ thứ gì, gõ ra bất cứ thứ gì, cũng như mở và chạy bất kỳ phần mềm nào chúng tìm thấy – chẳng hạn như mở một trình duyệt web và nghịch ngợm trên mạng.

Nếu nhìn vào tương lai của việc học sâu tăng cường và các cải tiến sau đó, thì chúng ta không thể thấy rõ đích đến nào. Tiềm năng cho ứng dụng này không chỉ giới hạn trong thế giới trò chơi ảo, vì nếu bạn là rô-bốt thì chính sự sống cũng có thể được xem như một trò chơi. Stuart Russell từng nói với tôi rằng, khoảnh khắc “quỷ thần ơ” lớn lao đầu tiên đến khi anh nhìn con rô-bốt Big Dog chạy lên sườn dốc phủ đầy tuyết trong rừng; nó đã nhẹ nhàng giải quyết vấn đề về chuyển động chân mà bản thân anh đã phải vật lộn để xử lý trong nhiều năm.³ Để đạt cột mốc này vào năm 2008, các lập trình viên đã phải khôn khéo xử lý một lượng lớn công việc. Tuy nhiên, sau đột phá của DeepMind, không có điều gì cấm một con rô-bốt chọn một ứng dụng học sâu tăng cường để tự học đi bộ mà không cần các lập trình viên: nó chỉ cần một hệ thống cho điểm mỗi khi nó tiến bộ. Tương tự, rô-bốt trong thế giới thực cũng có thể học bơi lội, bay, chơi bóng bàn, chiến đấu và gần như vô tận các nhiệm vụ vận động khác mà không cần các lập trình viên con người giúp đỡ. Để tăng tốc độ, giảm nguy cơ bị mắc kẹt hay gây tổn hại cho bản thân trong quá trình học tập, chúng có vẻ sẽ cần thực hiện việc học này trong thế giới thực tế ảo ở giai đoạn đầu.

Trực giác, Sáng tạo và Chiến lược

Một thời khắc có ý nghĩa quyết định khác đối với tôi là khi hệ thống AI AlphaGo của DeepMind thắng trận đấu Cờ vây 5 ván với Lee Sedol, người được suy tôn là kỳ thủ hàng đầu thế giới vào đầu thế kỷ XXI.

Rất nhiều người cho rằng sẽ có lúc máy móc hạ bệ con người trong môn cờ vây; việc này đã từng xảy ra với các đồng nghiệp của

họ ở môn cờ vua hai thập kỷ trước. Tuy nhiên, vì hầu hết các chuyên gia cờ vây dự đoán xa hơn một thập kỷ nữa, nên chiến thắng của AlphaGo là khoảnh khắc quyết định đối với họ cũng như với tôi. Nick Bostrom và Ray Kurzweil cũng đã nhấn mạnh khó khăn trong việc tiên lượng đột phá của AI. Chúng ta cũng thấy rõ điều này từ cuộc phỏng vấn với chính Lee Sedol trước và sau khi thua ba ván đầu tiên:

Tháng 10 năm 2015: “Dựa trên trình độ từng thấy của nó. . . tôi nghĩ tôi sẽ thắng gần như tuyệt đối.”

Tháng 2 năm 2016: “Tôi nghe nói AI của Google DeepMind mạnh đến kinh ngạc và đang ngày càng mạnh hơn, nhưng tôi tự tin mình ít nhất có thể thắng lần này.”

Ngày 9 tháng 3 năm 2016: “Tôi rất ngạc nhiên vì tôi không nghĩ mình sẽ thua.”

Ngày 10 tháng 3 năm 2016: “Tôi thực không nói nên lời. . . Tôi đang thấy sốc. Phải thừa nhận rằng... ván thứ ba sẽ chẳng dễ dàng gì cho tôi cả. ”

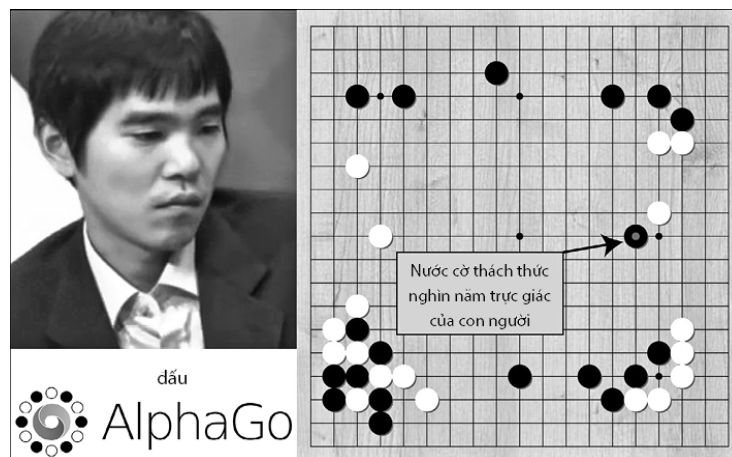
Ngày 12 tháng 3 năm 2016: “Tôi cảm thấy khá bất lực”.

Trong vòng một năm sau khi đấu với Lee Sedol, một AlphaGo cải tiến mạnh hơn đã đấu với 20 kỳ thủ hàng đầu thế giới mà không thua dù chỉ một trận.

Tại sao tôi lại coi đây là một vấn đề lớn với cá nhân tôi? Như tôi đã nói ở trên: trực giác và sáng tạo là hai đặc điểm cốt lõi của con người tôi. Bây giờ, tôi sẽ giải thích lý do tôi nghĩ AlphaGo bộc lộ cả hai điều này.

Trong môn cờ vây các kỳ thủ thay phiên nhau đặt quân đen và trắng trên bàn cờ 19x19 ô (xem hình 3.2). Có vô cùng nhiều thế cờ khả thi, hơn cả số nguyên tử có trong Vũ Trụ của chúng ta; bởi vậy phân tích tất cả hệ quả từ các nước cờ tương lai là điều vô ích. Do đó, các kỳ thủ chủ yếu dựa vào trực giác tiềm thức để bổ sung cho lý luận ý thức của họ, còn các cao thủ lại hình thành linh cảm gần như huyền bí về các thế cờ mạnh và yếu. Như đã thấy trong chương trước, kết quả của học sâu dường như rất giống trực giác: một mạng nơ-ron sâu có thể xác định một bức hình chứa một con mèo mà

không thể giải thích tại sao. Cũng vì lẽ này, nhóm nghiên cứu DeepMind đã cược rằng phương thức học sâu có thể nhận ra không chỉ những con mèo, mà còn cả những thế cờ vây mạnh. Họ đã xây dựng AlphaGo với ý tưởng cốt lõi là kết hợp sức mạnh trực giác của học sâu với sức mạnh lý luận của GOFAL – viết tắt của cụm từ hài hước “Good Old-Fashioned AI” (AI Thời Xưa Cũ) – được ứng dụng từ trước cuộc cách mạng học sâu. Họ đã sử dụng một cơ sở dữ liệu khổng lồ gồm các thế cờ vây trong cả ván đấu của con người lẫn các ván AlphaGo tự chơi với bản sao của chính nó, và đào tạo một mạng lưới nơ-ron sâu có thể dự đoán xác suất giúp quân trắng giành chiến thắng chung cuộc từ mỗi thế cờ. Họ cũng đào tạo một mạng lưới riêng biệt để dự đoán nước cờ tiếp theo. Sau đó, họ kết hợp các mạng này bằng phương pháp GOFAL, nó khéo léo tìm kiếm trong một danh sách các loạt nước cờ khả thi được trau chuốt nhằm xác định nước tiếp theo mang lại thế cờ chung cuộc mạnh nhất.



Hình 3.2: AI AlphaGo của DeepMind đã đi nước cờ rất sáng tạo ở hàng 5, thách thức hàng nghìn năm trí tuệ con người; khoảng 50 nước cờ sau đó đã chứng tỏ tính quyết định của nước cờ này trong việc đánh bại huyền thoại cờ vây Lee Sedol.

Sự kết hợp giữa trực giác và logic này đã sinh ra những nước cờ không chỉ mạnh mẽ, mà đôi khi còn rất sáng tạo. Ví dụ, trí tuệ đúc kết từ hàng nghìn năm lịch sử cờ vây cho rằng khi mới bắt đầu ván đấu, tốt nhất chỉ nên chơi ở hàng thứ ba hoặc thứ tư từ bên rìa. Hai cách này đều có điểm hay mà chúng ta cần lựa chọn: nếu chơi ở hàng thứ

ba, ta sẽ chiếm được lợi thế ngắn hạn là kéo lãnh thổ tới cạnh bàn cờ; nếu chơi ở hàng thứ tư, ta sẽ thu được lợi thế chiến lược lâu dài ở trung tâm.

Ở nước cờ thứ 37 của ván thứ hai, AlphaGo đã gây sốc cho giới cờ vây khi thách thức trí tuệ cổ xưa bằng nước cờ ở hàng thứ năm (hình 3.2), như thể nó còn tự tin hơn con người trong khả năng lập kế hoạch dài hạn và do đó ưa thích lợi thế chiến lược hơn cái lợi ngắn hạn. Các nhà bình luận đều choáng váng; Lee Sedol thậm chí phải đứng dậy và tạm thời rời phòng đấu.⁴ Như thể đã đoán trước, khoảng 50 nước cờ sau đó, cuộc chiến từ góc dưới bên trái của bàn cờ đã tràn qua và kết nối với quân đen ở nước 37 đó! Lối chơi này đã dẫn tới chiến thắng chung cuộc của AlphaGo, tôn vinh nước cờ hàng thứ năm của AlphaGo như một trong những nước cờ sáng tạo nhất lịch sử cờ vây.

Bởi cờ vây yêu cầu trực giác và sáng tạo nên nó thường được coi là một hình thức nghệ thuật hơn là trò chơi. Người ta đánh giá cờ vây là một trong bốn loại hình nghệ thuật tinh túy của Trung Quốc cổ đại, cùng với vẽ tranh, thư pháp và nhạc cổ cầm, và ngày nay nó vẫn rất phổ biến ở châu Á, với gần 300 triệu người xem trận đấu đầu tiên giữa AlphaGo và Lee Sedol. Bởi lẽ đó, giới cờ vây đã rúng động vì kết quả này và xem chiến thắng của AlphaGo như một cột mốc sâu sắc đối với nhân loại. Kỳ thủ cờ vây hàng đầu thế giới ở thời điểm đó là Kha Khiết có nói:⁵ “Nhân loại đã chơi cờ vây hàng ngàn năm, vậy mà AI đã cho thấy chúng ta còn chưa chạm tới được bề mặt của bộ môn này... Sự kết hợp giữa kỳ thủ con người và máy tính sẽ mở ra một kỷ nguyên mới... Cùng nhau, con người và AI sẽ tìm ra chân lý của cờ vây.” Sự hợp tác thành công giữa con người và máy tính như vậy đầy hứa hẹn ở nhiều lĩnh vực, bao gồm khoa học, nơi chúng ta hy vọng AI sẽ giúp con người hiểu biết sâu sắc hơn và nhận ra tiềm năng tối thượng của mình.

Đối với tôi, AlphaGo còn cho chúng ta một bài học quan trọng khác trong tương lai gần: logic của GOF AI kết hợp với trực giác của học sâu có thể tạo ra các chiến lược vô song. Vì cờ vây là một trong những trò chơi chiến thuật tối thượng, nên AI hiện đang trong tư thế sẵn sàng vươn lên và thách thức (hoặc giúp đỡ) những chiến lược

gia giỏi nhất ở các lĩnh vực khác ngoài trò chơi – ví dụ như trong các chiến lược đầu tư, chính trị hay quân sự. Những vấn đề chiến lược thực tế đó thường rất phức tạp do ảnh hưởng tâm lý con người, thiếu thông tin và các yếu tố cần được mô phỏng ngẫu nhiên, nhưng hệ thống AI chơi bài poker đã chứng minh rằng không có thách thức nào trong số đó là không thể vượt qua.

Ngôn ngữ tự nhiên

Một lĩnh vực khác mà gần đây AI đã tiến bộ đến mức khiến tôi sửng sốt là ngôn ngữ. Từ thời trẻ tôi vốn đã mê du lịch, và lòng hiếu kỳ đối với các nền văn hóa và ngôn ngữ khác đã trở thành một phần bản tính quan trọng trong tôi. Tôi lớn lên trong môi trường tiếng Thụy Điển và tiếng Anh, được dạy tiếng Đức và tiếng Tây Ban Nha ở trường, học tiếng Bồ Đào Nha và Rumani qua hai cuộc hôn nhân và tự học tiếng Nga, tiếng Pháp và tiếng Quan Thoại cho vui.

But the AI has been reaching, and after an important discovery in 2016, there are almost no lazy languages that I can translate between better than the system of the AI developed by the equipment of the brain of Google.

(Nhưng AI đang vươn tới, và sau một phát hiện quan trọng vào năm 2016, hầu như không có ngôn ngữ lười biếng nào mà tôi có thể dịch xuôi ngược tốt hơn so với hệ thống AI được thiết bị não của Google phát triển.)

Tôi đã thể hiện rõ ý của mình chưa nhỉ? Thực ra tôi định nói thế này:

But AI has been catching up with me, and after a major breakthrough in 2016, there are almost no languages left that I can translate between better than the AI system developed by the Google Brain team.

(Nhưng AI đã bắt kịp tôi, và sau một bước đột phá lớn vào năm 2016, hầu như không còn ngôn ngữ nào mà tôi có thể dịch xuôi ngược tốt hơn hệ thống AI được nhóm Google Brain phát triển.)

Tuy nhiên, ban đầu tôi đã dịch câu này sang tiếng Tây Ban Nha và ngược lại bằng một ứng dụng trên máy tính xách tay vài năm trước. Vào năm 2016, nhóm Google Brain đã nâng cấp dịch vụ

Google Dịch miễn phí của họ bằng một mạng các nơ-ron hồi quy sâu, giúp cải thiện mạnh mẽ các hệ thống GOFAL cũ.⁶

But AI has been catching up on me, and after a breakthrough in 2016, there are almost no languages left that can translate between better than the AI system developed by the Google Brain team.

(Nhưng AI đã bắt kịp tôi, và sau một bước đột phá lớn vào năm 2016, hầu như không còn ngôn ngữ nào có thể dịch xuôi ngược tốt hơn hệ thống AI được nhóm Google Brain phát triển.)

Như bạn có thể thấy, chúng ta đã lạc mất đại từ “tôi” khi dịch vòng sang tiếng Tây Ban Nha và không may thay, điều này đã thay đổi ý nghĩa câu. Đúng, nhưng chưa chuẩn! Tuy nhiên, để biện minh cho AI của Google, dù sao thì người ta vẫn thường chỉ trích tôi là hay viết văn lủng củng và khó phân tích, và tôi đã chọn một trong những câu loằng ngoằng khó hiểu nhất của mình cho ví dụ này. Đối với các câu thường dùng hơn, AI của họ vẫn dịch hoàn hảo. Độ chính xác này đã khiến hàng trăm triệu người dùng xôn xao vì nó đủ hữu ích để sử dụng hằng ngày. Hơn nữa, nhờ những tiến bộ gần đây trong học sâu ở lĩnh vực chuyển đổi từ giọng nói sang văn bản và ngược lại, những người dùng này giờ đây có thể nói với điện thoại thông minh của họ bằng một ngôn ngữ và nghe kết quả đã được dịch sang ngôn ngữ họ muốn.

Xử lý ngôn ngữ tự nhiên hiện là một trong những lĩnh vực tiến bộ nhanh nhất của AI và tôi nghĩ rằng thành công đó sẽ tác động rất lớn tới chúng ta, bởi ngôn ngữ là một đặc điểm trọng tâm của con người. AI dự đoán ngôn ngữ càng tốt thì sẽ càng giỏi trong việc soạn email trả lời hoặc tiếp tục một cuộc trò chuyện. Chỉ ít với người ngoài cuộc, điều này tạo cảm giác có sự xuất hiện của tư duy con người. Do đó, hệ thống học sâu đang chập chững tiến tới vượt qua phép thử Turing nổi tiếng, ở đó một cỗ máy phải trò chuyện đủ tốt bằng văn bản để đánh lừa con người rằng nó cũng là con người.

Tuy nhiên, AI xử lý ngôn ngữ vẫn cần đi một chặng đường khá dài. Tuy tôi thừa nhận bản thân cảm thấy hơi nhụt chí khi AI dịch tốt hơn, nhưng tôi lại thấy khá hơn mỗi khi tự nhắc mình rằng, cho đến nay nó vẫn không hiểu ý nghĩa của những gì nó nói. Được đào tạo dựa trên các tập dữ liệu khổng lồ, nó chỉ phát hiện ra các khuôn mẫu

và mối quan hệ liên quan đến từ ngữ, chứ không bao giờ liên hệ những từ này với bất kỳ thứ gì trong thế giới thực. Ví dụ, AI gán mỗi từ với một danh sách hàng nghìn con số định rõ sự tương đồng của từ đó với một số từ nhất định. Sau đó, nó có thể kết luận rằng sự khác biệt giữa vua và hoàng hậu tương tự như giữa chồng và vợ – nhưng nó vẫn không biết chút nào về ý nghĩa của việc là nam hay nữ, thậm chí không biết là có một thực thể như vậy ở thế giới thực với giá trị không gian, thời gian và vật chất.

Vì phép thử Turing có đặc thù là đánh lừa, nên người ta hay phê phán rằng nó giống một bài kiểm tra tính ngây thơ của con người chứ không phải một bài kiểm tra trí tuệ nhân tạo thực sự. Ngược lại, một phép thử đối trọng mang tên Winograd Schema Challenge (Thách thức Giản đồ Winograd) tấn công trực tiếp vào nhược điểm đó, nhằm thách thức hiểu biết thường thức mà hệ thống học sâu hiện tại thiếu sót. Con người chúng ta thường xuyên sử dụng kiến thức thực tế khi phân tích một câu, để xem đại từ đề cập đến danh từ nào. Ví dụ, một thách thức Winograd điển hình sẽ hỏi “họ” ở đây chỉ thứ gì:

1. “Hội đồng thành phố từ chối trao giấy phép cho những người biểu tình vì họ sợ bạo lực.”
2. “Hội đồng thành phố từ chối trao giấy phép cho những người biểu tình vì họ ủng hộ bạo lực.”

Có một cuộc thi hàng năm cho các AI trả lời các câu hỏi như vậy và các AI vẫn thể hiện khá tệ.⁷ Chính thách thức từ ngữ này đã khiến cả Google Dịch tê liệt khi tôi thay thế tiếng Tây Ban Nha bằng tiếng Trung trong ví dụ trên:

But the AI has caught up with me, after a major break in 2016, with almost no language, I could translate the AI system than developed by the Google Brain team.

(Nhưng AI đã bắt kịp với tôi, sau một kỳ nghỉ lớn vào năm 2016, với hầu như không có ngôn ngữ nào, tôi có thể dịch hệ thống AI hơn thứ được nhóm Google Brain phát triển.)

Hãy tự mình thử tại trang <https://translate.google.com> ngay bây giờ khi đang đọc sách này và xem liệu AI của Google đã được cải thiện chưa! Có khả năng nó đã được cải thiện, vì có nhiều cách tiếp

cận triển vọng cho cuộc hôn nhân giữa mạng nơ-ron hồi quy sâu với GOFAL nhằm xây dựng một AI xử lý ngôn ngữ có chứa mô hình về thế giới.

Cơ hội và Thách thức

Ba ví dụ này rõ ràng chỉ là mẩu thử, bởi AI đang tăng tốc nhanh chóng trên nhiều mặt trận quan trọng. Hơn nữa, mặc dù tôi chỉ đề cập đến hai công ty trong các ví dụ này, các nhóm nghiên cứu cạnh tranh nhau tại các trường đại học và các công ty khác cũng không mấy tụt hậu. Ta có thể nghe văng vẳng những lời mời hào phóng đến từ Apple, Baidu, DeepMind, Facebook, Google, Microsoft và các công ty khác, như thể họ đang dùng một cỗ máy hút sạch về họ các sinh viên, nghiên cứu sinh tiến sỹ và giảng viên các khoa viện khoa học máy tính trên khắp thế giới.

Đừng để các ví dụ tôi đưa ra đánh lừa mà xem lịch sử AI là các giai đoạn trì trệ điểm xuất phát đột phá. Trái lại, theo quan điểm của tôi, AI đã có quá trình tiến bộ khá ổn định trong một thời gian dài – nhưng truyền thông vẫn hay dùng từ “bước đột phá” bất cứ khi nào AI bước tới ngưỡng đánh dấu một ứng dụng thu hút trí tưởng tượng hoặc một sản phẩm hữu ích mới. Do đó, tôi đánh giá khá khả quan rằng AI sẽ tiếp tục phát triển nhanh trong nhiều năm tới. Hơn nữa, như chúng ta đã thấy trong chương trước, không có lý do cơ bản nào mà tiến trình này không thể tiếp tục cho đến khi khả năng của AI ngang ngửa con người ở hầu hết các nhiệm vụ.

Điều này đặt ra câu hỏi: Việc đó sẽ tác động đến chúng ta như thế nào? Tiến triển AI trong ngắn hạn sẽ thay đổi cách sống của con người chúng ta sống ra sao? Chúng ta đã thấy ngày càng khó có thể tranh luận rằng AI hoàn toàn không có mục tiêu, hiểu biết rộng, trực giác, sức sáng tạo hoặc ngôn ngữ – vốn là những đặc điểm mà nhiều người cho là tiền đề cho việc làm người. Điều này có nghĩa là ngay cả trong ngắn hạn, trước cả khi một AGI bất kỳ có thể ngang bằng chúng ta ở mọi nhiệm vụ, AI có thể ảnh hưởng đáng kể tới cách chúng ta nhìn nhận bản thân, tới những gì chúng ta có thể làm khi được AI hỗ trợ và tới cách chúng ta có thể kiếm sống khi cạnh tranh với AI. Tác động này sẽ khiến mọi việc tốt hơn hay xấu hơn? Điều này sẽ tạo ra những cơ hội và thách thức nào trong tương lai gần?

Mọi điều chúng ta yêu thích về nền văn minh đều là sản phẩm của trí tuệ con người, vì vậy, nếu có thể khuếch đại nó với trí tuệ nhân tạo, rõ ràng chúng ta sẽ có khả năng khiến cuộc sống tốt hơn. Ngay cả những tiến bộ khiêm tốn trong AI cũng có thể chuyển thành những cải tiến quan trọng trong khoa học công nghệ và sự giảm thiểu tai nạn, bệnh tật, bất công, chiến tranh, lao động nặng và nghèo đói. Nhưng để không tạo ra thêm vấn đề mới trong khi nhận lấy lợi ích từ AI, chúng ta cần trả lời nhiều câu hỏi quan trọng. Chẳng hạn:

1. Chúng ta có thể làm gì để hệ thống AI trong tương lai ổn định hơn hiện tại, để chúng thực hiện mong muốn của chúng ta mà không bị treo, hỏng hóc hoặc bị tấn công?

2. Chúng ta có thể làm gì để cập nhật hệ thống pháp lý của mình khiến nó công bằng, hiệu quả hơn và có thể bắt kịp với môi trường kỹ thuật số thay đổi nhanh chóng?

3. Chúng ta có thể làm gì để tạo ra vũ khí thông minh hơn và ít khả năng tấn công dân thường vô tội hơn mà không chầm ngòi cho một cuộc đua ngoài tầm kiểm soát về vũ khí tự hành có thể gây chết người?

4. Chúng ta có thể làm gì để phát triển thịnh vượng nhờ tự động hóa mà không phát sinh tình trạng thiếu thu nhập hoặc mất việc?

Chúng ta sẽ dành phần còn lại của chương này để khám phá lần lượt từng câu hỏi trên. Bốn câu hỏi về ngắn hạn này chủ yếu lần lượt nhắm tới các nhà khoa học máy tính, học giả pháp lý, các nhà chiến lược quân sự và các nhà kinh tế học. Tuy nhiên, để có được câu trả lời cần thiết vào đúng lúc cần, tất cả mọi người cần tham gia vào cuộc đối thoại này. Bởi như chúng ta sẽ thấy, những thách thức này vượt qua mọi ranh giới truyền thống – qua cả các lĩnh vực ngành nghề cũng như quốc gia.

AI ổn định và các Lỗi

Công nghệ thông tin đã và đang có tác động tích cực và lớn lao đối với mọi lĩnh vực hoạt động con người: từ khoa học đến tài chính, sản xuất, giao thông vận tải, y tế, năng lượng và truyền thông. Tuy vậy, tác động này chẳng là gì so với tiềm năng mà tiến bộ AI có thể mang lại. Nhưng chúng ta càng dựa vào công nghệ, thì công nghệ

càng cần ổn định, đáng tin cậy và có thể thực hiện những gì chúng ta muốn.

Trong suốt lịch sử nhân loại, chúng ta vẫn chỉ dựa vào một phương pháp hiệu quả để đảm bảo công nghệ luôn hữu ích: học hỏi từ những sai lầm. Chúng ta phát minh ra lửa, liên tục gây chuyện với nó, rồi sau đó mới phát minh ra bình cứu hỏa, lối thoát hiểm, thiết bị báo cháy và sở cứu hỏa. Chúng ta phát minh ra ô tô, liên tục đâm xe, rồi sau đó mới phát minh ra dây an toàn, túi khí và xe tự lái. Cho đến nay, tác động tiêu cực của các công nghệ của chúng ta là vừa đủ thấp để các điểm lợi vẫn vượt qua các điểm hại. Tuy nhiên, vì chúng ta sẽ không ngừng phát triển các công nghệ ngày càng mạnh hơn, chắc chắn sẽ tới lúc dù chỉ một tai nạn thôi cũng đủ để đánh đổ lợi ích nó đem lại. Một số người tranh luận rằng chiến tranh hạt nhân toàn cầu ngẫu nhiên xảy ra sẽ là một ví dụ. Những người khác lại cho rằng sẽ có một đại dịch từ kỹ thuật sinh học. Trong chương tiếp theo, chúng ta sẽ khám phá những tranh cãi xung quanh việc liệu AI trong tương lai có khiến con người tuyệt chủng hay không. Nhưng chúng ta không cần đến những ví dụ cực đoan như vậy để đưa ra một kết luận quan trọng: khi công nghệ thêm mạnh mẽ, chúng ta nên bớt dựa vào phương pháp thử-và-sai để thiết kế an toàn. Nói cách khác, chúng ta nên chủ động hơn là bị động, đầu tư vào nghiên cứu an toàn để ngăn ngừa hoàn toàn nguy cơ tai nạn. Đây cũng chính là lý do xã hội đầu tư vào điện hạt nhân an toàn nhiều hơn là bẫy chuột an toàn.

Và như chúng ta đã thấy trong chương 1, đây cũng là lý do cộng đồng tại hội nghị ở Puerto Rico lại quan tâm về nghiên cứu an toàn AI đến vậy. Tuy máy tính và hệ thống AI vẫn hay gặp sự cố, nhưng lần này lại khác: AI đang dần bước vào thế giới thực và nếu nó phá tan lưới điện, thị trường chứng khoán hay hệ thống vũ khí hạt nhân thì chuyện sẽ không chỉ dừng lại ở phiền toái. Trong đoạn còn lại của phần này, tôi muốn giới thiệu với bạn bốn lĩnh vực chính trong nghiên cứu kỹ thuật an toàn AI, hiện đang là trọng tâm của các cuộc thảo luận an toàn AI trên toàn thế giới: kiểm định, thẩm định, bảo an và kiểm soát.²⁵ Để mọi thứ không chán ngắt và khô khan, chúng ta sẽ khám phá những thành công và thất bại của công nghệ thông tin

thuộc các lĩnh vực khác nhau trong quá khứ, cũng như các bài học quý giá và các nghiên cứu thách thức từ những bài học này.

Tuy chỉ đề cập đến các hệ thống máy tính lạc hậu mà chẳng mấy ai xem là AI và gây ra ít thương vong (nếu có), nhưng những câu chuyện xưa cũ này vẫn sẽ dạy chúng ta những bài học quý giá trong việc thiết kế hệ thống AI an toàn và mạnh mẽ trong tương lai. Nếu những hệ thống tương lai này gặp lỗi, chúng có thể sẽ gây ra thảm họa thực sự.

AI thăm dò vũ trụ

Chúng ta hãy bắt đầu với một điều thân thương với tôi: khám phá vũ trụ. Công nghệ máy tính đã giúp chúng ta đưa con người lên Mặt trăng và gửi tàu vũ trụ không người lái khám phá tất cả các hành tinh trong Hệ Mặt Trời, thậm chí đổ bộ lên sao chổi và vệ tinh Titan của Sao Thổ. Trong chương 6, chúng ta sẽ thấy AI trong tương lai có thể giúp chúng ta khám phá các hệ mặt trời và thiên hà khác – ấy là nếu nó không gặp lỗi. Ngày 4 tháng 6 năm 1996, các nhà khoa học đang mong chờ nghiên cứu từ quyển Trái đất đã được reo hò hân hoan khi tên lửa Ariane 5 của Cơ quan Vũ trụ châu Âu gầm rú phóng lên không trung cùng các thiết bị khoa học mà họ chế tạo. Tuy nhiên, nụ cười của họ vụt tắt sau 37 giây khi tên lửa phát nổ, gây nên một màn trình diễn pháo hoa trị giá hàng trăm triệu đô-la.⁸ Nguyên nhân được cho là do phần mềm lỗi đã đưa vào một con số quá lớn, vượt quá 16 bit đơn vị được phân bổ cho nó.⁹ Hai năm sau, Mars Climate Orbiter (Tàu Quỹ đạo thăm dò khí hậu Sao Hỏa) của NASA vô tình thâm nhập vào bầu khí quyển của Hành Tinh Đỏ và vỡ tan tành, vì hai phần khác nhau của phần mềm sử dụng đơn vị đo lực khác nhau, gây ra sai số 445% ở thiết bị điều khiển lực đẩy tên lửa.¹⁰ Đây là lỗi siêu đắt giá thứ hai của NASA: ngày 22 tháng 7 năm 1962, tàu Mariner 1 thực hiện sứ mệnh tới Sao Kim đã phát nổ sau khi cất cánh ở Mũi Canaveral vì phần mềm điều khiển chuyển bay ngừng hoạt động do một dấu câu không chính xác.¹¹ Như để chứng minh không chỉ phương Tây làm chủ nghệ thuật phóng lỗi sai bay vào vũ trụ, sứ mệnh tàu Phobos 1 của Liên Xô cũng thất bại vào ngày 2 tháng 9 năm 1988. Đây là phi thuyền liên hành tinh nặng nhất từng được phóng, với mục tiêu phi thường là đổ bộ một tàu lên vệ tinh Phobos

của Sao Hỏa. Sứ mệnh này tan tành khi mã lệnh gửi đi thiếu mất một dấu gạch ngang, khiến lệnh “kết thúc sứ mệnh”, tắt tất cả hệ thống của phi thuyền trong khi nó còn đang trên đường đến Sao Hỏa.¹²

Từ những ví dụ này, chúng ta học được tầm quan trọng của bước kiểm định với các nhà khoa học máy tính: đảm bảo phần mềm đáp ứng được đầy đủ mọi yêu cầu dự kiến. Càng nhiều sự sống và tài nguyên bị đe dọa, chúng ta càng phải chắc chắn rằng phần mềm sẽ hoạt động như đã định. May mắn thay, AI có thể giúp tự động hóa và cải thiện quá trình xác minh. Ví dụ, gần đây một nhân hệ điều hành đa dụng hoàn chỉnh mang tên seL4 đã được xác minh toán học thông qua một đặc tả hình thức để đảm bảo không gặp sự cố và vận hành an toàn. Tuy chưa đi kèm nhiều tiện ích phụ hấp dẫn như Microsoft Windows và Mac OS, nhưng bạn có thể yên tâm rằng nó sẽ không đem đến những hiện tượng như màn hình xanh chết chóc hay bánh xe diệt vong quay vòng.²⁶ Cơ Quan Nghiên Cứu Dự Án Quốc Phòng Tiên Tiến (DARPA) của Hoa Kỳ đã tài trợ cho việc phát triển bộ các công cụ đảm bảo cao cấp dùng mã nguồn mở HACMS (hệ thống quân sự đảm bảo cao cấp)²⁷ có thể chứng minh độ tin cậy. Thách thức quan trọng là làm cho công cụ như vậy đủ mạnh mẽ và dễ sử dụng để có thể triển khai rộng rãi. Thách thức khác là bản thân nhiệm vụ kiểm định sẽ trở nên khó khăn hơn khi phần mềm chuyển sang dạng rô-bốt và môi trường mới, và khi phần mềm lập trình sẵn truyền thống bị thay thế bằng hệ thống AI có khả năng liên tục học tập, qua đó thay đổi hành vi của chúng, như trong chương 2.

AI cho tài chính

Tài chính cũng là một lĩnh vực được công nghệ thông tin chuyển hóa, giúp tái phân bổ các nguồn lực trên toàn cầu hiệu quả hơn với tốc độ ánh sáng. Công nghệ thông tin cũng khiến việc cấp vốn dễ dàng hơn cho mọi đối tượng, từ các khoản thế chấp tới vốn hỗ trợ công ty khởi nghiệp. AI tiến bộ cũng sẽ mang lại cơ hội lợi nhuận lớn trong tương lai từ giao dịch tài chính: hầu hết các quyết định mua/bán trên thị trường chứng khoán giờ đây được máy tính thực hiện tự động và mức lương khởi điểm cực khủng cho công việc cải thiện thuật toán giao dịch vẫn thường lôi cuốn các sinh viên của tôi tại MIT.

Kiểm định cũng là bước quan trọng với phần mềm tài chính; đây chính là điều mà hãng Knight Capital của Mỹ đã học được qua kinh nghiệm đau thương ngày 1 tháng 8 năm 2012, khi mất 440 triệu đô-la sau 45 phút triển khai phần mềm giao dịch chưa được kiểm định.¹³ Vụ “Sụp đổ chóng vánh” nghìn tỷ đô-la xảy ra ngày 6 tháng 5 năm 2010 lại gây chú ý vì một lý do khác. Mặc dù nó gây gián đoạn thị trường trên diện rộng trong khoảng nửa giờ trước khi ổn định lại, khiến giá cổ phiếu của một số công ty nổi tiếng như Procter & Gamble dao động giữa một xu và 100.000 đô-la¹⁴, vấn đề lại không nằm ở lỗi hoặc trục trặc máy tính có thể phòng tránh qua kiểm định. Thay vào đó, vấn đề xảy ra do giả định bị vượt quá: lúc ấy, các chương trình giao dịch tự động từ nhiều công ty đang hoạt động trong một tình huống bất ngờ, ở đó giả định của chúng không hợp lệ – ví dụ như giả định khi một máy tính trao đổi chứng khoán báo cáo rằng một cổ phiếu có giá một xu thì cổ phiếu đó thực sự đáng giá một xu.

Vụ sụp đổ chóng vánh này minh họa tầm quan trọng của thẩm định với các nhà khoa học máy tính: nếu kiểm định đặt ra câu hỏi “Tôi có xây dựng hệ thống chuẩn không?” thì thẩm định hỏi “Tôi có xây dựng chuẩn hệ thống không?”²⁸ Ví dụ, liệu hệ thống có dựa vào các giả định không phải lúc nào cũng hợp lệ? Nếu vậy, ta phải cải thiện thế nào để nó xử lý tính bất định này tốt hơn?

AI cho sản xuất

Dĩ nhiên, AI có tiềm năng rất lớn trong việc cải thiện sản xuất. Nó thực hiện điều này bằng cách kiểm soát các rô-bốt giúp tăng cường hiệu quả và độ chính xác. Máy in 3-D không ngừng cải tiến nay đã có thể tạo ra nguyên mẫu của bất kỳ thứ gì, từ các tòa nhà văn phòng đến thiết bị cơ học có kích thước nhỏ hơn một hạt muối.¹⁵ Trong khi các rô-bốt công nghiệp hạng nặng lắp ráp ô tô và máy bay, thì máy tính lại điều khiển các loại máy cán, tiện, cắt quy mô nhỏ hơn để cung cấp nguồn lực không chỉ cho các nhà máy, mà còn cả “phong trào sáng chế” cấp cơ sở. Phong trào này gồm những người đam mê ở địa phương; họ hiện thực hóa ý tưởng tại hơn 1.000 phòng chế tạo cộng đồng trên toàn thế giới.¹⁶ Nhưng khi càng có nhiều rô-bốt ở xung quanh chúng ta, việc chúng ta kiểm định và thẩm định phần

mềm của chúng càng trở nên quan trọng. Robert Williams là người đầu tiên bị rô-bốt giết. Anh là công nhân tại nhà máy của Ford ở Flat Rock, Michigan. Năm 1979, một rô-bốt có nhiệm vụ lấy các linh kiện ở khu vực cất giữ bị trục trặc, nên anh đã leo vào khu vực đó để tự mình lấy linh kiện ra. Con rô-bốt hoạt động trong yên lặng và đập vỡ đầu anh, liên tục trong 30 phút cho đến khi các đồng nghiệp của anh phát hiện ra sự vụ.¹⁷ Nạn nhân tiếp theo của rô-bốt là Kenji Urada, kỹ sư bảo dưỡng tại nhà máy Kawasaki ở Akashi, Nhật Bản. Trong lúc bảo dưỡng một rô-bốt bị hỏng vào năm 1981, anh vô tình bật công tắc của nó và bị cánh tay rô-bốt thủy lực nghiền chết.¹⁸ Năm 2015, một đại diện nhà thầu 22 tuổi tại một nhà máy của Volkswagen ở Baunatal, Đức, đang cài đặt cho một rô-bốt lấy các phụ tùng ô tô và thao tác với chúng. Một sự cố đã xảy ra, khiến rô-bốt túm lấy và nghiền anh lên một tấm kim loại đến chết.¹⁹

Mặc dù những tai nạn này rất bi thảm, nhưng ta cũng cần lưu ý là chúng chỉ chiếm một phần cực nhỏ trong tổng số tai nạn công nghiệp. Hơn nữa, tai nạn công nghiệp đã thay đổi tỷ lệ nghịch với phát triển công nghệ, giảm từ khoảng 14.000 ca tử vong năm 1970 xuống chỉ còn 4.821 vào năm 2014 tại Hoa Kỳ.²⁰ Ba tai nạn kể trên cho thấy, bổ sung trí tuệ cho những cỗ máy vốn vô tri sẽ cải thiện an toàn công nghiệp hơn nữa, bằng cách dạy cho rô-bốt cần cẩn thận hơn khi ở quanh con người. Cả ba vụ tai nạn này sẽ không xảy ra nếu quy trình thẩm định tốt hơn, vì các rô-bốt này gây hại không phải do lỗi hoặc ác ý mà bởi chúng đã đưa ra giả định không hợp lệ – rằng không ai có mặt ở đó hoặc rằng nạn nhân là phụ tùng ô tô.

Hình 3.3: Trong khi các rô-bốt công nghiệp truyền thống rất đắt và khó lập trình, xu hướng hiện nay là ứng dụng các rô-bốt rẻ hơn điều khiển bằng AI, có khả năng học những gì cần làm từ các công nhân không có kinh nghiệm lập trình.

AI cho giao thông vận tải

AI có khả năng bảo vệ nhiều sinh mạng trong sản xuất công nghiệp, nhưng nó có khả năng bảo vệ còn nhiều hơn thế trong giao thông vận tải. Trong năm 2015, chỉ riêng tai nạn xe hơi đã cướp đi 1,2 triệu tính mạng; tai nạn máy bay, xe lửa và thuyền bè cộng lại cũng đã giết chết hàng ngàn người khác. Ngay tại Hoa Kỳ, nơi có tiêu

chuẩn an toàn cao, tai nạn ở các xe có động cơ cũng đã gây ra cái chết cho 35.000 người trong năm 2016 – gấp bảy lần tổng số vụ tai nạn công nghiệp.²¹ Khi chúng tôi thảo luận về vấn đề này hồi năm 2016 trong cuộc họp thường niên của Hiệp hội vì sự Tiến bộ của Trí tuệ nhân tạo tại Austin, Texas, ông Moshe Vardi, một nhà khoa học máy tính Israel, tỏ ra khá khẩn thiết và lập luận rằng AI không nên chỉ có khả năng giảm tai nạn xe cộ, mà nó buộc phải vậy. Ông thốt lên: “Đó là một thôi thúc đạo đức!” Bởi hầu như tất cả các tai nạn xe hơi đều do lỗi của con người, nên nhiều người tin rằng xe ô tô tự lái do AI vận hành có thể giảm ít nhất 90% số người chết trên đường và sự lạc quan này đang thúc đẩy mạnh mẽ tiến trình thực sự đưa những chiếc xe tự lái ra đường. Elon Musk thì hình dung rằng những chiếc xe tự lái trong tương lai không chỉ an toàn hơn mà còn kiếm tiền cho chủ sở hữu của chúng khi chúng không được dùng đến, bằng cách cạnh tranh với các dịch vụ như Uber và Lyft.

Cho đến nay, xe tự lái rõ ràng vẫn an toàn hơn hẳn so với xe người lái, và các tai nạn xảy ra đã nhấn mạnh tầm quan trọng và khó khăn của bước thử nghiệm. Ngày 14 tháng 2 năm 2016, ô tô tự lái của Google gây ra vụ va quệt nhỏ đầu tiên do giả định sai về xe buýt: rằng tài xế xe buýt trên xe sẽ nhường đường khi nó chạy ra. Vụ tai nạn chết người đầu tiên xảy ra ngày 7 tháng 5 năm 2016, khi một xe tự lái Tesla đâm vào một xe container đang trên đường cao tốc. Tai nạn này xảy ra do hai giả định tệ:²² rằng phần sườn màu trắng sáng của toa xe cũng là một phần của bầu trời chói chang và rằng người lái xe (nghe nói đang bận xem một tập phim Harry Potter) sẽ luôn chú ý và can thiệp nếu có điều không ổn.²⁹

Nhưng đôi khi, kiểm định và thử nghiệm tốt cũng chưa đủ để tránh tai nạn, vì chúng ta cần cả kiểm soát tốt, tức cần có người giám sát được một hệ thống và thay đổi được hành vi của hệ thống này nếu cần. Để hệ thống bán tự hành này hoạt động tốt, ta rất cần một giao thức hiệu quả giữa người và máy. Ví dụ hiệu quả ở đây là ánh đèn đỏ trên nắp ô tô sẽ cảnh báo được cho bạn thấy cốp xe đang để mở. Ngược lại, ngày 6 tháng 3 năm 1987 khi chiếc phà chở tàu xe Herald of Free Enterprise của Anh rời Cảng Zeebrugge với cửa trước mũi phà vẫn mở toang, thuyền trưởng chẳng nhận được cảnh

báo ánh sáng hay hình ảnh nào. Phà lật không lâu sau khi rời cảng, đoạt mạng 193 người.²³

Một tai nạn bi thảm khác do kiểm soát có thể tránh khỏi nếu có giao thức hiệu quả giữa người và máy xảy ra vào đêm 01 tháng 6 năm 2009, khi máy bay Air France 447 lao xuống Đại Tây Dương, khiến tất cả 228 người trên chuyến bay thiệt mạng. Theo báo cáo chính thức, “phi hành đoàn lúc ấy không phát hiện ra động cơ máy bay đã ngừng nên không thực hiện thủ thuật khởi động lại” – bao gồm việc cho mũi máy bay chúc xuống – cho đến khi quá muộn. Các chuyên gia an toàn bay cho rằng, vụ tai nạn này đáng lý không xảy ra nếu có đồng hồ báo “góc tấn” trong buồng lái, để các phi công thấy mũi máy bay đã hướng lên trên quá cao.²⁴

Ngày 20 tháng 1 năm 1992, khi chuyến bay số hiệu 148 của hãng Air Inter đâm vào dãy núi Vosges gần Strasbourg ở Pháp khiến 87 người thiệt mạng, nguyên nhân không phải do thiếu giao thức máy-người mà do giao diện người dùng khó hiểu. Các phi công đã nhập dãy “33” trên bàn phím vì họ muốn chúc xuống 3,3 độ, nhưng phần mềm tự lái diễn giải nó thành 3,300 feet mỗi phút vì hệ thống đang ở chế độ khác – màn hình hiển thị lại quá nhỏ nên các phi công không thể kịp thời nhìn ra chế độ vận hành và nhận ra sai lầm của họ.

AI cho năng lượng

Công nghệ thông tin đã làm nên nhiều điều kỳ diệu trong sản xuất và phân phối điện, với các thuật toán tinh vi cân bằng giữa sản xuất và tiêu thụ của các lưới điện thế giới, cùng những hệ thống kiểm soát tinh vi giúp các nhà máy điện hoạt động an toàn và hiệu quả. Những tiến bộ của AI trong tương lai có thể làm cho “lưới điện thông minh” còn thông minh hơn để thích nghi tối đa với mức độ cung- cầu luôn thay đổi, đến tận quy mô từng tấm pin mặt trời trên mái nhà và hệ thống pin gia đình. Nhưng vào Thứ Năm ngày 14 tháng 8 năm 2003, khoảng 55 triệu người ở Hoa Kỳ và Canada rơi vào tình trạng mất điện, trong đó rất nhiều người không được cấp điện suốt nhiều ngày sau. Nguyên nhân chính ở đây cũng do lỗi trong giao thức máy-người: một lỗi phần mềm đã khiến hệ thống báo động trong một trạm điều khiển tại Ohio không cảnh báo nhân viên vận hành về nhu cầu

phân phối lại nguồn điện trước khi một vấn đề nhỏ (dây cao thế trần quá tải chạm vào tán lá um tùm) bùng ra ngoài tầm kiểm soát.²⁵

Vụ nóng chảy một phần lõi lò hạt nhân tại lò phản ứng trên đảo Three Mile ở Pennsylvania vào ngày 28 tháng 3 năm 1979 đã tiêu tốn khoảng một tỷ đô-la chi phí dọn dẹp và khiến người dân phản ứng rất dữ dội với điện hạt nhân. Báo cáo tổng kết đã xác định nhiều yếu tố góp phần, bao gồm cả sự nhầm lẫn do giao diện người dùng kém.²⁶ Cụ thể, cảnh báo ánh sáng mà nhân viên vận hành tưởng rằng biểu thị một van an toàn quan trọng được mở hay đóng thực tế chỉ cho biết tín hiệu đóng van đã được gửi đến hay chưa – vì vậy các nhân viên vận hành đã không nhận ra rằng van này hở.

Những tai nạn về năng lượng và giao thông vận tải này đã cho chúng ta thấy, khi để AI phụ trách thêm nhiều hệ thống vật lý, chúng ta cần nỗ lực nghiên cứu nghiêm túc - để các cỗ máy không chỉ tự hoạt động tốt mà còn cộng tác hiệu quả với những người kiểm soát. Khi AI thông minh hơn, điều này sẽ không chỉ gói gọn ở việc xây dựng giao diện người dùng tốt cho việc chia sẻ thông tin nữa, mà còn bao gồm tìm ra cách phân bổ nhiệm vụ tối ưu trong nhóm làm việc máy-người – ví dụ như cách xác định các tình huống cần chuyển giao quyền kiểm soát cũng như áp dụng hiệu quả cách đánh giá của con người vào những quyết định có giá trị nhất, thay vì khiến những người kiểm soát mất tập trung bởi hàng loạt thông tin không quan trọng.

AI cho y tế

AI cũng có tiềm năng lớn trong việc cải thiện ngành y tế. Việc số hóa bệnh án vẫn đang giúp các bác sĩ và bệnh nhân ra quyết định nhanh hơn, tốt hơn và nhận được trợ giúp tức thì từ các chuyên gia chẩn đoán hình ảnh kỹ thuật số trên khắp thế giới. Rõ ràng, với các tiến bộ nhanh chóng trong lĩnh vực thị giác máy tính và học sâu, các hệ thống AI sẽ sớm trở thành những chuyên gia giỏi nhất trong chẩn đoán hình ảnh. Ví dụ, một nghiên cứu của Hà Lan vào năm 2015 cho thấy máy tính chẩn đoán về ung thư tiền liệt tuyến sử dụng ảnh chụp cộng hưởng từ (MRI) cũng chính xác như bác sĩ X-quang,²⁷ và nghiên cứu ở Stanford vào năm 2016 cho thấy AI có thể chẩn đoán ung thư phổi bằng hình ảnh hiển vi thậm chí còn chính xác hơn cả

các nhà bệnh lý học.²⁸ Nếu học máy có thể giúp tiết lộ mối quan hệ giữa gen, bệnh tật và cách cơ thể phản ứng với điều trị, nó có thể đem đến một cuộc cách mạng trong y tế cá nhân, khiến các loại gia súc gia cầm trở nên khỏe mạnh hơn và tạo ra cây trồng có sức chống chịu tốt hơn. Hơn nữa, rô-bốt còn có tiềm năng trở thành bác sĩ phẫu thuật chính xác và đáng tin cậy hơn cả con người, ngay cả khi không sử dụng AI tiên tiến. Trong những năm gần đây, rô-bốt đã thực hiện thành công hàng loạt ca phẫu thuật; các thủ thuật này thường rất chính xác, sử dụng các công cụ nhỏ cùng những vết mổ nhỏ hơn giúp giảm mất máu, giảm đau và giảm thời gian hồi phục.

Nhưng buồn thay, ngành công nghiệp y tế cũng đã có những bài học đau đớn về tầm quan trọng của phần mềm ổn định. Ví dụ, máy xạ trị Therac-25 do Canada sản xuất có hai chế độ điều trị bệnh nhân ung thư khác nhau: sử dụng hoặc chùm electron năng lượng thấp, hoặc chùm tia X năng lượng cao có hiệu điện thế hàng triệu volt mạnh được một chiếc khiên đặc biệt hội tụ vào mục tiêu. Thật không may, phần mềm chưa kiểm định và còn lỗi của máy đã khiến các kỹ thuật viên chiếu chùm tia mạnh không kèm khiên hội tụ thay vì chiếu chùm năng lượng thấp. Nhiều bệnh nhân đã thiệt mạng.²⁹ Nhiều bệnh nhân tại Viện Ung thư Quốc gia ở Panama cũng đã tử vong do chiếu xạ quá liều: thiết bị xạ trị ở đây sử dụng phóng xạ từ đồng vị cobalt-60 đã bị lập trình thời gian chiếu xạ quá dài vào năm 2000 và 2001 do giao diện người dùng khó hiểu mà chưa được kiểm định hợp lệ.³⁰ Cũng theo một báo cáo gần đây, từ năm 2000 đến năm 2013,³¹ tai nạn phẫu thuật rô-bốt có liên quan tới 144 ca tử vong và 1.391 người bị thương ở Mỹ, với các vấn đề thường gặp bao gồm không chỉ các lỗi phần cứng như chập hồ quang điện làm các mảnh cháy hoặc mảnh vỡ rơi vào bệnh nhân, mà còn cả lỗi phần mềm như chuyển động không kiểm soát hay tắt nguồn đột ngột.

Tin tốt lành là phần còn lại của gần hai triệu ca phẫu thuật rô-bốt trong phạm vi báo cáo diễn ra suôn sẻ và rô-bốt dường như đã giúp thực hiện phẫu thuật an toàn hơn. Theo một nghiên cứu của chính phủ Hoa Kỳ, chất lượng chăm sóc y tế kém đã góp phần gây ra hơn 100.000 ca tử vong mỗi năm chỉ riêng ở Hoa Kỳ,³² vì vậy việc phát

triển AI cho y học thậm chí còn cấp thiết về mặt đạo đức hơn so với phát triển AI cho xe tự lái.

AI cho truyền thông

Công nghiệp truyền thông là ngành mà máy tính có tác động lớn hơn cả từ trước đến nay. Kể từ khi các tổng đài điện thoại vi tính xuất hiện vào những năm 1950, Internet vào những năm 1960, và World Wide Web³⁰ vào năm 1989, ngày nay, hàng tỷ người lên mạng để giao tiếp, mua sắm, đọc tin tức, xem phim hoặc chơi trò chơi; họ đã quen thuộc với thế giới mà thông tin chỉ cách họ một cú nhấp chuột – và thường miễn phí. Sự trỗi dậy của Internet vạn vật cũng hứa hẹn nâng cao hiệu quả, độ chính xác, sự thuận tiện và lợi ích kinh tế qua việc đưa mọi thứ lên mạng, bao gồm từ đèn đóm, bộ ổn nhiệt thay tủ lạnh đến các mạch thu phát sinh học trên gia súc gia cầm.

Những thành công ngoạn mục này trong việc kết nối thế giới đã mang tới cho các nhà khoa học máy tính thách thức thứ tư: họ cần cải thiện không chỉ quy trình kiểm định, thẩm định và kiểm soát mà cả về mặt bảo an để chống lại phần mềm độc hại và tin tặc tấn công. Trong khi những vấn đề đã đề cập ở các phần trước đều là hệ quả từ những sai lầm không chủ đích, thì tính bảo an lại nhắm tới hành động phi pháp có chủ đích. Phần mềm độc hại đầu tiên thu hút mạnh mẽ sự chú ý của giới truyền thông là phần mềm sâu Morris, được tung ra vào ngày 2 tháng 11 năm 1988, khai thác các lỗi trong hệ điều hành UNIX. Nghe nói đó là một nỗ lực sai lầm nhằm đếm số lượng máy tính đang trực tuyến, và mặc dù nó đã lây lan và làm sập khoảng 10% trong 60.000 máy tính tạo nên Internet hồi đó, nhưng việc này cũng không ngăn nổi MIT bổ nhiệm người viết ra nó là ông Robert Morris vào vị trí giảng dạy chính thức tại khoa khoa học máy tính.

Những phần mềm độc hại khác thì không khai thác lỗ hổng trong phần mềm mà ở con người. Ngày 5 tháng 5 năm 2000, như thể để kỷ niệm sinh nhật của tôi, nhiều người nhận được email với tiêu đề “ILOVEYOU” từ người quen và đồng nghiệp, và những người dùng hệ điều hành Microsoft Windows nào đã nhấp chuột vào tệp đính kèm “LOVE-LETTER-FOR-YOU.txt.vbs” sẽ vô tình khởi động một tập lệnh làm hỏng máy tính và gửi lại email đó cho mọi người trong sổ địa chỉ của họ. Được tạo ra bởi hai lập trình viên trẻ ở Philippines,

con sâu máy tính này đã lây lan cho khoảng 10% mạng Internet, tương tự sâu Morris, nhưng vì Internet lúc này đã lớn hơn rất nhiều lần, nên nó đã trở thành một trong những vụ lây lan lớn nhất mọi thời đại, ảnh hưởng tới hơn 50 triệu máy tính và gây ra thiệt hại hơn 5 tỷ đô-la. Hẳn bạn cũng nhớ biết, rằng Internet vẫn bị nhiễm vô số loại phần mềm độc hại có tính lây nhiễm mà các chuyên gia bảo mật phân loại thành sâu, Trojan, virus và các nhóm với tên gọi đáng sợ khác, và chúng gây ra những thiệt hại từ việc phát các thông điệp chơi xỏ vô hại trên máy, xóa các tệp của bạn, ăn cắp thông tin cá nhân hay theo dõi bạn cho đến khống chế máy tính của bạn để gửi thư rác.

Trong khi phần mềm độc hại nhắm tới bất kỳ máy tính nào có thể, các tin tặc lại tấn công các mục tiêu rất cụ thể — các ví dụ nổi cộm gần đây bao gồm Target, TJ Maxx, Sony Pictures, Ashley Madison, công ty dầu khí Aramco của Ả Rập và Ủy ban Quốc gia Dân chủ Hoa Kỳ. Hơn nữa, lượng tài sản bị mất dường như ngày càng ngoạn mục hơn. Tin tặc đã đánh cắp 130 triệu số thẻ tín dụng và các thông tin tài khoản khác từ Hệ thống Thanh toán Heartland trong năm 2008, và xâm phạm hơn một tỷ(!) tài khoản email Yahoo! trong năm 2013.³³ Một vụ tin tặc tấn công năm 2014 vào Văn phòng Quản lý Cán bộ Chính phủ Hoa Kỳ đã xâm phạm hồ sơ nhân sự và thông tin hồ sơ xin việc của hơn 21 triệu người; người ta cho là những người bị ảnh hưởng bao gồm cả các công chức với quyền tiếp cận thông tin mật tối cao và dấu vân tay của đặc vụ ngầm.

Vì lẽ đó, tôi thường ngờ vực mỗi khi đọc về những hệ thống mới được suy tôn là an toàn và bất khả xâm phạm 100%. Tuy nhiên, “bất khả xâm phạm” lại chính là điều chúng ta cần ở hệ thống AI tương lai trước khi chúng ta giao cho chúng trọng trách, trong những việc như cơ sở hạ tầng hoặc hệ thống vũ khí tối quan trọng chẳng hạn ; chính vì lẽ đó, nếu vai trò của AI trong xã hội ngày càng lớn, thì tầm quan trọng của an ninh máy tính cần ngày một cao. Trong khi một số vụ tấn công khai thác tính cả tin của con người hoặc lỗi hổng phức tạp trong phần mềm mới phát hành, thì một số vụ lại giúp tin tặc đăng nhập trái phép vào máy tính từ xa bằng cách lợi dụng các lỗi đơn giản nhưng không được ai để ý trong một thời gian dài. Ta có lỗi hổng

“Heartbleed” kéo dài từ năm 2012 tới năm 2014 trong một trong những thư viện phần mềm phổ biến nhất về giao tiếp an toàn giữa các máy tính, còn lỗ hổng “Bashdoor” thì nằm sâu trong hệ điều hành Unix từ năm 1989 đến năm 2014. Điều này có nghĩa rằng các công cụ AI cải thiện kiểm định và thẩm định cũng sẽ cải thiện tính bảo an.

Thật không may, các hệ thống AI tốt hơn cũng có thể được sử dụng để tìm ra các lỗ hổng mới và thực hiện các vụ tấn công phức tạp hơn. Ví dụ, thử tưởng tượng một ngày bạn nhận được email “lừa đảo” được cá nhân hóa đến mức bất thường nhằm thuyết phục bạn tiết lộ thông tin cá nhân. Nó được gửi từ tài khoản một người bạn của bạn, bởi AI đã tấn công và mạo danh cô ấy, bắt chước phong cách viết của cô ấy qua phân tích các email khác đã được gửi và bao gồm nhiều thông tin cá nhân về bạn từ các nguồn khác. Liệu bạn có sập bẫy? Nếu như email lừa đảo có vẻ đến từ công ty phát hành thẻ tín dụng của bạn và nó đi kèm một cuộc điện thoại với giọng con người thân thiện mà bạn không thể xác định là do AI tạo nên thì sao? Trong cuộc chạy đua vũ trang về bảo mật máy tính đang diễn ra giữa bên tấn công và bên phòng thủ, có rất ít dấu hiệu cho thấy phe phòng thủ đang chiến thắng.

Luật pháp

Con người chúng ta là loài động vật xã hội, đã chinh phục tất cả các giống loài khác và thống trị Trái Đất nhờ khả năng hợp tác. Chúng ta đã phát triển luật để khuyến khích và thúc đẩy việc hợp tác, vì vậy nếu AI có thể cải thiện hệ thống pháp lý và quản trị của chúng ta, thì nó có thể giúp chúng ta hợp tác thành công hơn trước đây nhiều và phát huy hết khả năng. Và có rất nhiều cơ hội để cải thiện ở đây, cả về cách chúng ta xây dựng và áp dụng luật, nên ta sẽ lần lượt khám phá cả hai.

Khi bạn nghĩ tới hệ thống tòa án ở đất nước mình, những điều đầu tiên bạn nghĩ tới là gì? Nếu đó là sự chậm trễ kéo dài, chi phí cao và đôi lúc còn bất công, thì bạn không phải người duy nhất nghĩ vậy. Sẽ không tuyệt vời sao, nếu suy nghĩ đầu tiên của bạn là “hiệu quả” và “công bằng”? Do ta có thể xem quy trình pháp lý theo cách trừu tượng như một quá trình tính toán, nhập vào thông tin về bằng

chứng và luật lệ rồi đưa ra quyết định, một số học giả mơ về việc tự động hóa hoàn toàn ngành này bằng các rô-bốt-thẩm-phán: một hệ thống AI không ngừng áp dụng các tiêu chuẩn pháp lý cao cho mọi quyết định mà không mắc phải các sai sót con người do thiên vị, mệt mỏi hay thiếu cập nhật kiến thức.

Rô-bốt-thẩm-phán

Năm 1994, Byron De La Beckwith Con bị kết tội ám sát lãnh đạo quyền dân sự Medgar Evers vào năm 1963, nhưng hai bồi thẩm đoàn riêng rẽ toàn người da trắng ở Mississippi đã không thể kết án hắn ta ngay vào sau năm xảy ra vụ giết người, dù bằng chứng vật lý về cơ bản là giống nhau.³⁴ Đáng buồn thay, lịch sử pháp lý dường như đầy rẫy những phán xét thiên vị bị ảnh hưởng của màu da, giới tính, khuynh hướng tình dục, tôn giáo, quốc tịch và các yếu tố khác. Về nguyên tắc, rô-bốt-thẩm-phán có thể đảm bảo rằng lần đầu tiên trong lịch sử, mọi người thực sự bình đẳng trước pháp luật: chúng ta có thể lập trình để chúng giống hệt nhau và đối xử bình đẳng với mọi người, áp dụng luật theo cách minh bạch và hoàn toàn không thiên vị ai.

Rô-bốt-thẩm-phán cũng có thể loại bỏ những thành kiến ngẫu nhiên chứ không cố ý của con người. Ví dụ, một nghiên cứu gây tranh cãi năm 2012 kết luận rằng các thẩm phán Israel thường đưa ra phán quyết khắc nghiệt hơn hẳn khi đối: họ từ chối khoảng 35% trường hợp tạm tha ngay sau khi ăn sáng, và từ chối hơn 85% trường hợp ngay trước bữa trưa.³⁵

Một thiếu sót khác của các thẩm phán con người là họ có thể không đủ thời gian để khám phá tất cả các chi tiết của vụ án. Ngược lại, vì chỉ là phần mềm nên rô-bốt-thẩm-phán có thể dễ dàng nhân bảnhờ đó có thể xử lý song song tất cả các vụ án đang chờ chứ không cần tuần tự. Mỗi rô-bốt-thẩm-phán sẽ được chỉ định cho từng vụ án riêng mà không cần giới hạn thời gian. Cuối cùng, các thẩm phán con người không thể nắm vững tất cả các kiến thức kỹ thuật cần thiết cho mọi vụ án, từ các tranh chấp căng thẳng về bằng sáng chế tới các vụ giết người bí ẩn phụ thuộc hoàn toàn vào công nghệ pháp y tân thời nhất, nhưng các rô-bốt-thẩm-phán trong tương lai sẽ có bộ nhớ và khả năng học hỏi không giới hạn.

Tới một ngày, những rô-bốt-thẩm-phán như vậy có thể nhờ đó mà hiệu quả hơn và công bằng hơn, dựa vào tính không thiên vị, năng lực cao và hoạt động minh bạch. Tính hiệu quả của chúng càng làm cho chúng thêm công bằng: bằng cách đẩy nhanh tiến trình pháp lý và khiến các luật sư khôn ngoan khó bẻ cong kết luận hơn, chúng sẽ giúp việc tìm lại công lý thông qua các tòa án ít tốn kém hơn nhiều. Điều này có thể làm tăng đáng kể cơ hội của một cá nhân hoặc công ty khởi nghiệp thiếu tiền chống lại một tỷ phú hoặc tập đoàn đa quốc gia có cả một quân đoàn luật sư.

Mặt khác, nếu rô-bốt-thẩm-phán có lỗi hoặc bị tấn công thì sao? Cả hai thứ này đều từng tác động xấu tới các máy kiểm phiếu tự động, và khi phải cân đo giữa nhiều năm sau song sắt với hàng triệu đô-la trong ngân hàng, thì động cơ cho các cuộc tấn công mạng hãy còn đè nặng. Ngay cả khi AI đủ ổn định để chúng ta tin rằng một rô-bốt-thẩm-phán đang sử dụng thuật toán lập pháp phù hợp, liệu mọi người có hiểu lập luận của nó đủ để tôn trọng phán quyết nó đưa ra không? Thách thức này càng trở nên khó khăn hơn, bởi sự thành công gần đây của mạng thần kinh vốn khó hiểu nhưng lại thể hiện vượt trội hơn các thuật toán AI dễ hiểu thông thường. Nếu bị cáo muốn biết lý do họ bị kết án, hẳn họ có quyền nhận được câu trả lời tử tế hơn là “chúng tôi huấn luyện hệ thống bằng nhiều dữ liệu, và đây là phán quyết nó đưa ra” phải không? Hơn nữa, các nghiên cứu gần đây đã chỉ ra rằng nếu bạn đào tạo một hệ thống nơ-ron học sâu với một lượng dữ liệu tù nhân khổng lồ, nó có thể dự đoán ai có khả năng tái phạm (và từ chối tạm tha người này) tốt hơn các thẩm phán con người. Nhưng nếu hệ thống này phát hiện ra sự tái phạm có liên hệ về mặt thống kê với giới tính hoặc chủng tộc của tù nhân thì liệu điều đó có thể hiện rằng rô-bốt-thẩm-phán này phân biệt giới tính, phân biệt chủng tộc và cần được lập trình lại không? Thật vậy, một nghiên cứu năm 2016 lập luận rằng các phần mềm dự đoán tái phạm được ứng dụng khắp Hoa Kỳ đang có thiên hướng chống lại người Mỹ gốc Phi và đã góp phần tạo nên các bản án bất công.³⁶ Đây đều là những câu hỏi quan trọng mà tất cả chúng ta cần suy ngẫm và thảo luận để đảm bảo rằng AI vẫn giúp ích cho chúng ta. Chúng ta đang đối mặt với không phải một quyết định “được ăn cả, ngã về

không” về việc ứng dụng rô-bốt-thẩm-phán, mà là một quyết định về mức độ và tốc độ triển khai ứng dụng AI trong hệ thống pháp luật. Liệu chúng ta có muốn một hệ thống hỗ trợ quyết định dựa trên AI dành cho thẩm phán con người, giống như cho các bác sĩ trong tương lai không? Hay chúng ta muốn đi xa hơn và nhận các quyết định từ các rô-bốt-thẩm-phán, nhưng có thể kháng án sau với thẩm phán con người? Hay chúng ta muốn theo tới cùng và trao quyền tối thượng cho máy móc, ngay cả với án tử hình?

Tranh cãi về pháp lý

Tới đây, chúng ta mới chỉ khám phá việc áp dụng pháp luật; bây giờ chúng ta hãy chuyển sang nội dung của nó. Chúng ta vẫn đồng thuận rộng rãi rằng luật pháp của chúng ta cần phải phát triển để bắt kịp với công nghệ. Ví dụ, hai lập trình viên tạo ra sâu ILOVEYOU gây ra hàng tỷ đô-la thiệt hại ở trên được tòa tuyên trắng án và vẫn tự do vì tại thời điểm đó, ở Philippines chưa có luật chống lại việc tạo ra phần mềm độc hại. Vì tốc độ tiến bộ công nghệ dường như đang tăng nhanh, nên luật pháp cũng phải cập nhật nhanh hơn nhưng hiện tại lại đang có xu hướng cập nhật chậm trễ. Thu hút nhiều người hiểu biết về công nghệ vào các trường luật và chính phủ có lẽ là một động thái thông minh cho phát triển xã hội. Nhưng liệu sau đó ta có nên ứng dụng hệ thống hỗ trợ quyết định dựa trên AI cho cử tri và các nhà lập pháp, rồi tiếp nối sau đó là các nhà lập pháp rô-bốt?

Thay đổi luật pháp của chúng ta ra sao để theo kịp tiến bộ trong AI là chủ đề tranh cãi thú vị. Người ta tranh cãi căng thẳng giữa quyền riêng tư và tự do thông tin. Người theo hướng tự do lập luận rằng chúng ta càng có ít sự riêng tư hơn, thì tòa án sẽ càng có nhiều bằng chứng hơn, và phán xét sẽ công bằng hơn. Ví dụ: nếu chính phủ truy cập được thiết bị điện tử của mọi người để ghi lại nơi họ đang ở và những gì họ gõ, nhấp chuột, nói và làm, nhiều vụ án sẽ được giải quyết nhanh chóng và có thể ngăn chặn những vụ sau đó. Những người ủng hộ quyền riêng tư phản đối rằng họ không muốn một nhà nước giám sát mọi nơi mọi lúc, và kể cả nếu họ muốn, thì vẫn tồn tại nguy cơ nó sẽ biến thành một chế độ độc tài toàn trị với quy mô chưa từng có. Ngoài ra, các kỹ thuật học máy đã tiến bộ trong việc phân tích dữ liệu não từ máy quét fMRI³¹ để xác định được

một người đang nghĩ gì và, đặc biệt là, người đó nói thật hay nói dối.³⁷ Nếu công nghệ quét não được hỗ trợ bởi AI trở nên phổ biến trong các phòng xử án, thì quy trình đưa bằng chứng dài dòng có thể được đơn giản hóa đáng kể và diễn ra nhanh gọn, giúp các phiên tòa xét xử nhanh hơn và phán xét công bằng hơn. Nhưng phe ủng hộ quyền riêng tư lại lo lắng rằng liệu hệ thống như vậy đôi khi có mắc sai lầm và quan trọng hơn cả, liệu trí não chúng ta có nên nằm ngoài tầm rình mò của chính phủ hay không. Các chính phủ không ủng hộ tự do tư tưởng có thể sử dụng công nghệ này để hình sự hóa việc giữ gìn những niềm tin và ý kiến nhất định. Bạn sẽ vạch ra ranh giới giữa công lý và quyền riêng tư, cũng như giữa việc bảo vệ xã hội và bảo vệ quyền tự do cá nhân ở đâu? Dù bạn vạch rõ ở điểm nào, liệu nó có dần đi hẳn theo hướng giảm quyền riêng tư để bù đắp cho thực tế là bằng chứng đang dần dễ bị giả mạo hơn? Ví dụ, một khi AI có năng lực tạo dựng video giả hoàn toàn giống thật về cảnh bạn phạm tội, liệu bạn có bầu chọn một hệ thống cho phép chính phủ theo dõi mọi người ở mọi nơi mọi lúc và có thể cung cấp cho bạn chứng cứ ngoại phạm đánh thép khi cần?

Một tranh luận hấp dẫn khác là có nên quản lý các nghiên cứu AI hay không, hoặc khái quát hơn là những nhà hoạch định chính sách nên khích lệ các nhà nghiên cứu AI như thế nào để tăng tối đa cơ hội tạo ra kết quả hữu ích. Một số nhà nghiên cứu AI phản đối mọi hình thức quản lý việc phát triển AI, cho rằng các hình thức quản lý này sẽ trì hoãn những cải tiến cấp thiết (ví dụ, các xe tự lái nhằm giảm thiểu tai nạn) và sẽ đưa nghiên cứu AI tiên tiến vào vòng bí mật và/hoặc tới các quốc gia khác có chính phủ ưu đãi hơn. Tại hội nghị AI-có-ích tại Puerto Rico được đề cập ở chương đầu, Elon Musk cho rằng điều chúng ta cần ngay bây giờ từ chính phủ không phải là sự giám sát mà là thấu hiểu: cụ thể, cần bổ nhiệm những người có năng lực về mặt kỹ thuật vào chính quyền để theo dõi tiến trình AI và chỉ đạo nếu cần thiết. Anh cũng lập luận rằng quy định của chính phủ đôi khi có thể nuôi dưỡng chứ không trì hoãn sự tiến bộ: ví dụ, nếu tiêu chuẩn an toàn của chính phủ đối với xe tự lái giúp giảm số vụ tai nạn xe hơi tự lái, thì cộng đồng sẽ ít khả năng phản ứng dữ dội hơn và việc chấp nhận công nghệ mới sẽ nhanh chóng hơn. Do đó, các công ty

AI có ý thức an toàn cao nhất sẽ ủng hộ những chính sách quản lý bắt buộc các đối thủ cạnh tranh không tận tâm như họ phải đạt tới cùng tiêu chuẩn an toàn cao.

Một tranh cãi pháp lý thú vị khác lại liên quan đến việc cấp quyền cho máy móc. Nếu xe ô tô tự lái giúp giảm một nửa số lượng 32.000 người chết vì tai nạn giao thông hàng năm ở Mỹ, có lẽ các nhà sản xuất ô tô sẽ không nhận được 16.000 lời nhắn cảm ơn, mà là 16.000 vụ kiện. Vì thế, nếu một chiếc xe tự lái gây tai nạn, ai sẽ chịu trách nhiệm – người sử dụng, chủ sở hữu hay nhà sản xuất? Học giả pháp lý David Vladeck đã đề xuất phương án thứ tư: chính chiếc xe đó! Cụ thể, ông đề xuất rằng chiếc xe tự lái được phép (và bắt buộc phải) có bảo hiểm xe hơi. Bằng cách này, các mẫu xe có lịch sử an toàn xuất sắc sẽ chỉ phải trả rất ít phí bảo hiểm, có thể thấp hơn cả mức dành cho lái xe con người, trong khi các mẫu thiết kế kém từ các nhà sản xuất cầu thả sẽ chỉ được phép mua các gói bảo hiểm đắt đỏ, khiến chi phí cho chúng cao đến mức không ai thêm mua.

Nhưng nếu máy móc như xe hơi được phép có chính sách bảo hiểm, liệu chúng có nên sở hữu cả tiền và tài sản? Nếu vậy, sẽ chẳng còn điều luật nào ngăn các máy tính thông minh kiếm tiền trên thị trường chứng khoán và sử dụng nó để thanh toán cho các dịch vụ trực tuyến. Một khi máy tính có thể trả tiền để con người làm việc cho nó thì nó sẽ làm được bất cứ việc gì con người có thể làm. Việc hệ thống AI có thể giỏi hơn con người ở mặt đầu tư (điều mà chúng đã đạt được trong một số lĩnh vực), sẽ dẫn đến khả năng máy móc kiểm soát và sở hữu hầu hết nền kinh tế của chúng ta. Đây có phải là điều chúng ta muốn? Nếu nghe có vẻ xa vời, mong các bạn hãy nhớ rằng phần lớn nền kinh tế của chúng ta đã thuộc sở hữu của một thực thể phi nhân loại khác: các tập đoàn – thường có quyền lực lớn hơn bất kỳ ai bên trong chúng có sự sống riêng ở mức độ nào đó.

Nếu bạn chấp nhận việc cấp quyền sở hữu tài sản cho máy móc, thì có nên cho chúng cả quyền bỏ phiếu nữa không? Nếu có, liệu mỗi chương trình máy tính có nên được coi là một đầu phiếu hay không, mặc dù nó có thể dễ dàng tạo ra hàng tỷ tỷ bản sao của chính nó trong đám mây nếu nó đủ giàu và nhờ vậy đảm bảo quyền quyết định tất cả các cuộc bầu cử? Nếu không, chúng ta phân biệt đối xử giữa bộ não máy móc với bộ não con người trên cơ sở đạo đức nào? Nếu

bộ não máy móc có ý thức theo nghĩa là nó có trải nghiệm chủ quan như chúng ta, thì có điều gì khác biệt giữa người và máy không? Ta sẽ tìm hiểu sâu hơn những vấn đề gây tranh cãi liên quan đến tầm kiểm soát của máy tính đối với thế giới của chúng ta trong chương sau, và các vấn đề liên quan đến ý thức máy trong chương 8.

Vũ khí

Từ thời xa xưa không ai nhớ, loài người đã phải chịu cảnh đói nghèo, bệnh tật và chiến tranh. Chúng ta đã đề cập đến việc AI có thể giúp làm giảm nạn đói và bệnh tật, vậy còn chiến tranh thì sao? Một số người tranh luận rằng vũ khí hạt nhân ngăn chặn chiến tranh giữa những nước sở hữu vì chúng quá khủng khiếp, vậy tại sao không để tất cả các quốc gia chế tạo những vũ khí dựa trên AI còn khủng khiếp hơn với hy vọng đặt dấu chấm hết cho mọi cuộc chiến? Nếu bạn chưa thấy luận điểm này thuyết phục và tin rằng chiến tranh trong tương lai vẫn không thể tránh khỏi, vậy tại sao ta không sử dụng AI để khiến chiến tranh nhân đạo hơn? Nếu chiến tranh chỉ còn có máy móc đánh nhau với máy móc thì không cần quân nhân hay dân thường phải hy sinh tính mạng nữa. Ngoài ra, ta có thể hy vọng rằng các máy bay không người lái chạy bằng AI trong tương lai và các hệ thống vũ khí tự động (viết tắt là AWS, còn được phe phản đối gọi là “rô-bốt sát thủ”) sẽ công bằng và lý trí hơn binh lính con người: được trang bị các cảm biến siêu phàm và không sợ mất mạng, chúng có thể bình tĩnh tính toán và xử lý ngay cả trong lúc nước sôi lửa bỏng và ít có khả năng giết nhầm dân thường hơn.



Hình 3.4: Trong khi các máy bay không người lái ngày nay (như chiếc MQ-1 Predator này của Không quân Hoa Kỳ) là do con người điều khiển từ xa, các máy bay không người lái chạy bằng AI trong tương lai có tiềm năng loại bỏ hẳn con người khỏi việc vận hành, sử dụng thuật toán để quyết định ai sẽ là mục tiêu để tiêu diệt.

Con người tham gia vận hành

Nhưng nếu hệ thống tự động đó phát sinh lỗi, khó hiểu hay có hành vi không mong đợi thì sao? Hệ thống vũ khí đánh gần Phalanx của Hoa Kỳ dành cho tàu tuần dương lớp Aegis³² có thể tự động phát hiện, theo dõi và tấn công các mối đe dọa như tên lửa chống hạm và máy bay.

Tàu USS Vincennes là một tàu tuần dương mang tên lửa điều khiển có biệt danh là Robocruiser (tàu tuần dương rô-bốt) liên quan đến hệ thống Aegis trên tàu. Ngày 3/7/1988, trong một cuộc đụng độ với pháo hạm Iran trong chiến tranh Iran-Iraq, hệ thống radar của tàu đã cảnh báo về một máy bay đang tiếp cận. Thuyền trưởng William Rodgers III đã suy luận rằng một máy bay chiến đấu phản lực F-14 của Iran đang hạ độ cao và chuẩn bị tấn công họ nên đã ra lệnh cho phép hệ thống Aegis bắn. Lúc ấy, ông không hề nhận ra họ đã bắn rơi máy bay dân sự là chuyến bay số hiệu 655 của hãng hàng không Iran Air, khiến toàn bộ 290 người trên máy bay thiệt mạng và dư luận quốc tế phẫn nộ. Cuộc điều tra sau đó quy tội cho một giao diện người dùng khó hiểu, không tự động cho thấy chấm nào trên màn hình radar là máy bay dân sự (chuyến bay số hiệu 655 đi theo lộ trình bay thông thường hàng ngày và có bộ thu phát tín hiệu máy bay dân sự) hay phân biệt chấm nào đang hạ độ cao (để tấn công) và chấm nào đang nâng độ cao (lúc đó chuyến bay 655 đang cất cánh từ Tehran). Khi hệ thống tự động được yêu cầu cung cấp thông tin về chiếc máy bay bí ẩn, nó đã trả lời “đang hạ độ cao” vì quả thực có một máy bay đang ở trạng thái này, nhưng nó lại gán nhầm cho máy bay này số quản lý máy bay của hải quân Hoa Kỳ: thực ra máy bay đang hạ độ cao là một máy bay tuần tra không chiến mặt đất của Hoa Kỳ đang hoạt động ở Vịnh Oman cách khá xa nơi đó.

Trong ví dụ này, con người đã đưa ra quyết định cuối cùng. Do áp lực thời gian, vị thuyền trưởng đã quá tin vào hệ thống tự động. Theo các sĩ quan quốc phòng trên khắp thế giới, cho đến nay tất cả các hệ thống vũ khí tự động đều phụ thuộc quyết định con người, trừ những loại bẫy kỹ thuật thấp như mìn. Nhưng hiện nay người ta đang phát triển những vũ khí thực sự tự động, có thể lựa chọn và tấn công mục tiêu mà không cần đến con người. Về mặt quân sự, ý tưởng đưa toàn bộ con người ra khỏi vòng chiến để tăng tốc thật cảm dỗ: trong trận không chiến giữa một máy bay không người lái hoàn toàn tự động có thể phản ứng trong tích tắc với một máy bay không người lái phản ứng chậm hơn vì được điều khiển từ xa bởi một con người cách đó nửa vòng Trái Đất, bạn nghĩ bên nào sẽ thắng?

Tuy nhiên, đã có những tình huống ngàn cân treo sợi tóc mà chúng ta phải cho là cực kỳ may mắn khi có con người trong vòng chiến. Ngày 27/10/1962, trong cuộc Khủng hoảng tên lửa Cuba, 11 tàu khu trục Hải quân Hoa Kỳ và tàu sân bay USS Randolph đã bao vây tàu ngầm B-59 của Liên Xô ở gần Cuba, trong lãnh hải quốc tế bên ngoài khu vực “cách ly” của Hoa Kỳ. Họ không hay biết là nhiệt độ trên tàu đã tăng lên quá 45°C do nhiên liệu của tàu ngầm đang cạn và hệ thống điều hòa không khí đã ngừng hoạt động. Vì sắp ngộ độc cac-bon đi-ô-xit, nhiều thuyền viên đã ngất. Thủy thủ đoàn đã mất liên lạc với Moscow suốt nhiều ngày và không biết liệu Thế chiến III đã bắt đầu hay chưa. Rồi quân Mỹ bắt đầu thả những quả mìn nhỏ xuống: họ nói với Moscow là việc này chỉ để buộc tàu ngầm phải nổi lên và rời đi, nhưng thủy thủ đoàn không hề biết. “Chúng tôi nghĩ—đến lúc rồi—tận thế rồi”, thuyền viên V. P. Orlov nhớ lại. “Cứ như thể anh đang ngồi trong một cái thùng phuy, còn ở bên ngoài có người cứ lấy búa gõ vào thùng liên hồi vậy”. Một điều nữa mà quân Mỹ cũng không hay biết là thủy thủ đoàn B-59 có một ngư lôi hạt nhân mà họ được phép phóng mà không cần thông qua Moscow. Quả thực, lúc đó thuyền trưởng Savitski đã quyết tâm phóng ngư lôi hạt nhân. Valentin Grigorievich, sĩ quan phụ trách ngư lôi, đã quả quyết: “Quân ta sẽ chết, nhưng quân địch sẽ chìm hết—chúng ta sẽ không làm hổ danh hải quân Liên Xô!” Cũng rất may, quyết định phóng ngư lôi phải được ba sĩ quan trên tàu chấp thuận. Một trong ba người, Vasili Arkhipov, đã không đồng ý. Thật lạ là rất ít người biết đến

Arkhipov, mặc dù quyết định của ông có lẽ đã giúp nhân loại tránh khỏi Thế chiến III và có thể còn là đóng góp giá trị nhất trong lịch sử hiện đại của loài người.³⁸ Hẳn các bạn cũng không khỏi rung mình khi suy ngẫm về điều có thể xảy ra nếu B-59 là một tàu ngầm tự động do AI điều khiển mà không có con người tham gia vận hành.

Hai thập kỷ sau, vào ngày 9/9/1983, căng thẳng lại leo thang giữa hai siêu cường: Tổng thống Hoa Kỳ Ronald Reagan đã cáo buộc Liên bang Xô Viết là “đế chế ác quỷ”, và mới chỉ một tuần trước đó, Liên Xô đã bắn hạ một máy bay chở khách của hãng Hàng không Quốc gia Hàn Quốc đi lạc vào không phận của nước này, khiến 269 người thiệt mạng—trong đó có một nghị sĩ quốc hội Hoa Kỳ. Hôm ấy, một hệ thống cảnh báo sớm tự động hóa của Liên Xô báo cáo rằng Hoa Kỳ đã phóng năm tên lửa hạt nhân từ mặt đất nhắm tới Liên Xô, khiến sĩ quan Stanislav Petrov chỉ có vài phút để quyết định xem đây có phải báo động giả hay không. Vệ tinh được cho là đang hoạt động tốt, nên nếu theo đúng quy trình thì ông sẽ phải báo cáo về một vụ tấn công hạt nhân. Nhưng ông đã tin vào trực giác của mình, cho rằng Hoa Kỳ sẽ không tấn công chỉ với năm tên lửa nên đã báo cáo với tư lệnh cấp trên rằng đó là báo động giả mà không xác minh lại. Sau này, người ta mới biết là một vệ tinh đã nhầm lẫn các tia phản chiếu ánh sáng Mặt trời trên đỉnh các đám mây thành các tia phun ra từ động cơ tên lửa.³⁹ Tôi tự hỏi điều gì sẽ xảy ra nếu Petrov bị thay thế bởi một hệ thống AI chỉ biết làm theo đúng quy trình.

Đây có là cuộc chạy đua vũ trang tiếp theo?

Như bạn hẳn đã đoán được, cá nhân tôi có những lo ngại nghiêm túc về các hệ thống vũ khí tự động. Nhưng tôi vẫn chưa cho các bạn biết về nỗi lo chính của tôi: điểm kết thúc của cuộc chạy đua vũ khí AI. Tháng 7 năm 2015, tôi đã bày tỏ mối lo này trong bức thư ngỏ sau đây viết cùng Stuart Russell, với góp ý chân thành từ các đồng nghiệp tại Viện Tương lai Sự sống FLI:⁴⁰

VŨ KHÍ TỰ ĐỘNG:

Thư ngỏ từ các nhà nghiên cứu AI & Rô-bốt

Các loại vũ khí tự động lựa chọn và tấn công mục tiêu mà không cần con người can thiệp. Chẳng hạn, chúng gồm trực thăng bốn cánh quạt có vũ trang có khả năng tìm và diệt những người khớp với một số tiêu chí định sẵn, nhưng không bao gồm tên lửa hành trình hay máy bay không người lái điều khiển từ xa cần có con người để đưa ra quyết định ngắm bắn. Công nghệ Trí tuệ Nhân tạo (AI) nay đã đạt tới điểm mà chỉ trong vài năm tới chứ không cần đến vài thập kỷ tới là ta đã hoàn toàn có thể ứng dụng những hệ thống đó, kể cả về mặt pháp lý. Nguy cơ đến từ việc này rất cao: người ta cho rằng vũ khí tự động là một cuộc cách mạng thứ ba trong chiến tranh, chỉ sau thuốc súng và vũ khí hạt nhân.

Nhiều luận điểm đã được đưa ra nhằm ủng hộ hoặc chống lại vũ khí tự động, ví dụ như việc thay thế quân nhân bằng máy móc là có lợi vì sẽ giảm thương vong cho chủ thể, nhưng có hại vì sẽ hạ thấp rào cản cho yêu cầu tham chiến. Ngày nay, câu hỏi cốt yếu cho nhân loại là: liệu ta nên bắt đầu một cuộc chạy đua vũ trang AI toàn cầu hay ngăn chặn nó từ trong trứng nước. Nếu một cường quốc quân sự thúc đẩy phát triển vũ khí AI, chúng ta sẽ gần như không thể tránh khỏi một cuộc chạy đua vũ trang toàn cầu. Điểm kết thúc của con đường phát triển công nghệ này cũng đã rõ ràng: vũ khí tự động sẽ trở thành những khẩu AK của tương lai. Không như vũ khí hạt nhân, chúng không cần các thành phần đắt đỏ hay khó kiếm nên chúng sẽ có mặt ở khắp nơi với giá đủ rẻ để tất cả các cường quốc quân sự lớn có thể sản xuất hàng loạt. Sẽ chỉ là vấn đề thời gian cho tới khi chúng xuất hiện ở chợ đen và rơi vào tay những tên khủng bố, những kẻ độc tài muốn kiểm soát dân chúng triệt để hơn, những lãnh chúa quân đội muốn ra tay tàn sát để thanh lọc sắc tộc, v.v.. Vũ khí tự động sẽ là lựa chọn lý tưởng cho những nhiệm vụ như ám sát, gây bất ổn cho các quốc gia, đàn áp quần chúng và sát hại có chọn lọc một nhóm dân tộc cụ thể. Do đó, chúng tôi tin rằng một cuộc chạy đua vũ trang AI sẽ không có lợi cho nhân loại. Có nhiều cách để AI giúp chiến trường trở nên an toàn hơn cho con người, đặc biệt là dân thường, mà không cần phải tạo ra những công cụ giết người mới.

Cũng như hầu hết các nhà hóa học và nhà sinh học không có hứng thú với việc chế tạo vũ khí hóa học hay sinh học, hầu hết các nhà nghiên cứu AI cũng chẳng có hứng thú chế tạo vũ khí AI và

không muốn người khác gây ra vết nhơ cho ngành này bằng cách chế tạo chúng, vì có thể khiến công luận phản đối mạnh mẽ AI và hạn chế những lợi ích xã hội của AI trong tương lai. Quả thực, các nhà hóa học và sinh học đã ủng hộ mạnh mẽ những thỏa thuận quốc tế thành công trong việc cấm vũ khí hóa học và sinh học, cũng như hầu hết các nhà vật lý học ủng hộ các hiệp ước cấm vũ khí hạt nhân phóng từ không gian và các loại vũ khí laser gây mù mắt.

Để các lo ngại của chúng tôi khó bị làm ngơ như thể chúng tôi là một lũ gàn dở đấu tranh ôn hòa bảo vệ môi trường, tôi muốn càng nhiều nhà nghiên cứu AI và rô-bốt gao cội ký tên vào thư ngỏ của chúng tôi càng tốt. Chiến dịch Quốc tế Kiểm soát Vũ trang Rô-bốt³³ trước đó đã thu thập được hàng trăm chữ ký kêu gọi cấm rô-bốt sát thủ, và tôi cho là chúng tôi có thể làm tốt hơn thế. Tôi biết những tổ chức chuyên môn sẽ không muốn chia sẻ danh sách email thành viên khổng lồ của họ cho một mục đích có thể bị hiểu là mang tính chính trị, nên tôi đã góp nhặt danh sách các nhà nghiên cứu và khoa viện nơi họ làm việc từ các văn bản trên mạng và đăng quảng cáo về việc tìm kiếm địa chỉ email của họ trên MTurk—nền tảng cho đóng góp cộng đồng Amazon Mechanical Turk. Hầu hết các nhà nghiên cứu đều để địa chỉ email của họ trên trang web của trường đại học, và sau khi bỏ ra 24 giờ cùng 54 đô-la, tôi đã nghiễm nhiên có danh sách email của hàng trăm nhà nghiên cứu AI với thành tựu đủ để được bầu làm hội viên Hiệp hội Vì sự Tiến bộ của Trí tuệ Nhân tạo (AAAI³⁴). Một trong số họ là giáo sư AI người Úc gốc Anh Toby Walsh, người đã tốt bụng đồng ý email cho những người còn lại trong danh sách và giúp định hướng cho chiến dịch của chúng tôi. Những cộng tác viên MTurk khắp thế giới sau đó không ngừng gửi thêm các danh sách địa chỉ đến cho Toby, và không lâu sau, hơn 3000 nhà nghiên cứu AI và rô-bốt đã ký tên vào thư ngỏ của chúng tôi, bao gồm sáu cựu chủ tịch AAAI và các lãnh đạo ngành AI đến từ Google, Facebook, Microsoft and Tesla. Một đội quân các tình nguyện viên hùng hậu của FLI đã làm việc không mệt mỏi để xác minh danh sách chữ ký, loại bỏ những chữ ký lấy tên giả như Bill Clinton hay Sarah Connor³⁵. Hơn 17.000 người khác cũng ký tên, trong đó có Stephen Hawking. Sau khi Toby tổ chức một cuộc họp

báo về kết quả thu thập này tại Hội nghị Chung Quốc tế về Trí tuệ Nhân tạo³⁶, nó đã trở thành tin thời sự đưa đi khắp thế giới.

Vì các nhà sinh học và hóa học đã từng kiên quyết đấu tranh, nên giờ mọi người nhớ về các ngành đó vì đã tạo ra những loại thuốc và nguyên vật liệu có lợi thay vì vũ khí sinh học và hóa học. Giờ đây, cộng đồng AI và rô-bốt cũng đã lên tiếng: những người ký tên ủng hộ lá thư cũng muốn người khác nhớ đến ngành của họ vì đã tạo ra một tương lai tốt đẹp hơn, chứ không phải tạo ra những cách mới để giết người. Nhưng AI trong tương lai sẽ được dùng cho phục vụ mục đích dân sự hay quân sự? Mặc dù chúng ta đã dành nhiều trang trong chương này để bàn về các mục đích dân sự hơn, nhưng có thể chúng ta sẽ phải chi nhiều tiền hơn cho mục đích quân sự—đặc biệt nếu một cuộc chạy đua vũ trang AI bắt đầu. Các khoản đầu tư cho AI dân sự đã vượt quá một tỷ đô-la Mỹ vào năm 2016, nhưng con số này quá nhỏ so với yêu cầu ngân sách tài khóa 2017 của Lầu Năm Góc: vào khoảng 12 đến 15 tỷ đô-la Mỹ cho các dự án liên quan đến AI; Trung Quốc và Nga hẳn cũng lưu ý tuyên bố của Phó Bộ trưởng Quốc phòng Mỹ Robert Work, khi công bố ngân sách: “Tôi muốn các đối thủ của ta tự hỏi điều gì đang diễn ra sau tấm màn đen.”⁴¹

Nên có một hiệp ước quốc tế?

Mặc dù công luận quốc tế đang nỗ lực thúc đẩy việc đàm phán một lệnh cấm rô-bốt sát thủ, người ta vẫn chưa rõ điều gì sẽ xảy ra, và còn đang tranh luận sôi nổi rằng điều gì nên xảy ra. Tuy nhiên bên can dự hàng đầu nhất trí rằng các cường quốc nên soạn thảo một bộ luật quốc tế để định hướng nghiên cứu và ứng dụng AWS, nhưng họ vẫn chưa nhất trí về việc nên cấm cái gì và làm thế nào để thi hành lệnh cấm. Ví dụ, ta chỉ nên cấm vũ khí tự hành gây chết người hay cả những vũ khí gây thương tích nặng như gây mù mắt? Phải cấm việc phát triển, sản xuất hay sở hữu? Lệnh cấm có nên áp dụng cho tất cả các hệ thống vũ khí tự hành, hay như trong thư ngỏ của chúng tôi thì chỉ cấm những hệ thống tấn công, còn với những hệ thống phòng thủ như súng phòng không tự hành và hệ thống phòng thủ tên lửa thì cho phép? Trong trường hợp cho phép phòng thủ tên lửa, AWS có nên được tính là hệ thống phòng thủ dù chúng rất dễ triển khai trên lãnh thổ đối phương? Và làm cách nào để thi hành hiệp ước khi mà

hầu hết các bộ phận trong vũ khí tự vận hành cũng đồng thời có công dụng dân sự? Ví dụ, máy bay không người lái dùng để giao hàng cho Amazon với máy bay không người lái dùng để thả bom vốn chẳng mấy khác biệt.

Một số bên tranh luận đã lập luận rằng việc thành lập một hiệp ước về AWS hiệu quả là điều khó khăn không tưởng và do đó ta chẳng nên cố gắng cho chuyện này. Mặt khác, cố Tổng thống John F. Kennedy từng nhấn mạnh khi công bố các sứ mệnh lên Mặt trăng rằng những việc khó cũng rất đáng để thử nếu thành công của chúng có khả năng đem lại lợi ích to lớn cho tương lai nhân loại. Hơn nữa, nhiều chuyên gia tranh luận rằng các lệnh cấm vũ khí sinh học và hóa học đều có giá trị, dù việc thi hành chúng rất khó khăn với nhiều vụ lách luật, bởi các lệnh cấm này đã gây ra sự kỳ thị hóa nghiêm trọng khiến người ta hạn chế sử dụng chúng.

Tôi đã gặp Henry Kissinger tại một bữa tối trọng thể năm 2016 và có cơ hội hỏi về vai trò của ông trong lệnh cấm vũ khí sinh học. Ông kể, khi ông còn là Cố vấn an ninh quốc gia Hoa Kỳ, ông đã thuyết phục Tổng thống Nixon rằng lệnh cấm sẽ có lợi cho an ninh quốc gia Mỹ. Tôi rất ấn tượng về sự nhanh nhạy trong tư duy và trí nhớ của một người 92 tuổi như ông, và đã cuốn theo góc nhìn người trong cuộc của ông. Vì Hoa Kỳ vốn đã hưởng vị thế siêu cường nhờ thế mạnh quân sự hạt nhân và vũ khí thông thường nên nước này sẽ mất nhiều hơn được từ một cuộc chạy đua vũ trang sinh học toàn cầu với kết quả không chắc chắn. Nói cách khác, nếu bạn đã là kẻ đứng đầu, thì nên làm theo câu châm ngôn “nếu không hỏng thì đừng sửa”. Stuart Russell cũng tham gia cuộc đối thoại và chúng tôi cùng thảo luận về việc có thể áp dụng luận điểm tương tự khi nói về vũ khí tự hành chết người kẻ hưởng lợi nhiều nhất trong một cuộc chạy đua vũ trang không phải là các siêu cường, mà là các quốc gia tự cô lập nhỏ bé và các nhân tố phi quốc gia như khủng bố. Chính những nhân tố này sẽ được tiếp cận với vũ khí qua chợ đen sau khi chúng đã được phát triển hoàn thiện.

Một khi đã được sản xuất hàng loạt, chi phí sản xuất những máy bay AI sát thủ cỡ nhỏ có thể sẽ chỉ đắt hơn điện thoại thông minh một chút. Dù là một tên khủng bố muốn ám sát một chính trị gia hay một kẻ bị bạn gái phụ bạc muốn trả thù, họ chỉ cần tải hình ảnh và địa

chỉ của mục tiêu vào máy bay sát thủ: nó sẽ bay đến đích, xác định và tiêu diệt mục tiêu, rồi tự hủy để không ai có thể biết kẻ thủ ác. Còn nếu đó là những kẻ muốn thanh lọc sắc tộc thì máy bay có thể được lập trình để chỉ giết những người có màu da hay sắc tộc nhất định. Theo Stuart hình dung, những vũ khí đó càng thông minh thì lại càng cần ít nguyên liệu, hỏa lực và tiền bạc cho mỗi lần hạ sát. Ví dụ, anh e sợ một ngày sẽ có những máy bay không người lái có kích cỡ bằng con ong đất giết được người một cách rẻ mạt với hỏa lực rất yếu bằng cách bắn vào mắt, vì nơi này đủ mềm để một mảnh đạn nhỏ có thể xuyên thẳng vào não. Hoặc, chúng có thể bám vào đầu với những cặp càng kim loại và xuyên thủng sọ bằng một khối thuốc nổ siêu nhỏ. Nếu một triệu máy bay sát thủ như thế có thể ra quân từ trong thùng xe tải, thì người ta sẽ có một thứ vũ khí hủy diệt hàng loạt đáng sợ hoàn toàn mới: một loại vũ khí có thể tiêu diệt chỉ một loại người đã định mà không hề hấn gì đến tất cả những người hay những vật khác.

Một luận điểm đối nghịch lại cho rằng có thể loại bỏ những mối lo như vậy bằng cách làm cho các rô-bốt sát thủ trở nên quy củ—ví dụ, khiến chúng chỉ tiêu diệt địch. Nhưng nếu ta còn phải lo lắng về việc thực thi một lệnh cấm, thì làm thế nào mà việc cưỡng chế khiến vũ khí tự hành của đối phương phải quy củ 100% lại dễ dàng hơn việc cưỡng chế không sản xuất chúng ngay từ đầu? Và liệu có ai khẳng định được nhất quán rằng những binh lính được đào tạo bài bản của các quốc gia văn minh lại không thể tuân thủ những nguyên tắc chiến đấu mà rô-bốt có thể làm tốt hơn, trong khi đồng thời khẳng định những quốc gia tự cô lập, các độc tài và các nhóm khủng bố sẽ theo đúng nguyên tắc chiến tranh và không bao giờ triển khai các rô-bốt theo những cách có thể phạm vào các nguyên tắc đó?

Chiến tranh mạng

Một khía cạnh quân sự thú vị khác của AI là nó thông qua chiến tranh mạng có thể giúp bạn tấn công địch mà không cần chế tạo vũ khí. Như để báo trước những điều có thể trong tương lai, con sâu Stuxnet, được công luận cho là tác phẩm của chính phủ Mỹ và Israel, đã xâm nhập các máy ly tâm trong chương trình làm giàu hạt nhân của Iran và khiến chúng quay nhanh đến mức vỡ tan tành. Xã hội càng tự động hóa và AI tấn công càng mạnh mẽ hơn, thì chiến tranh

mạng càng gây ra thiệt hại lớn. Nếu bạn có thể thâm nhập và phá hoại những xe tự lái, máy bay tự lái, lò phản ứng hạt nhân, rô-bốt công nghiệp, hệ thống thông tin liên lạc, hệ thống tài chính và lưới điện của đối phương, thì bạn có thể hủy hoại hoàn toàn nền kinh tế và phá vỡ an ninh quốc phòng của họ. Nếu bạn có thể thâm nhập cả hệ thống vũ khí của họ nữa thì càng hay.

Chúng ta đã bắt đầu chương này bằng cách khảo sát các cơ trong tương lai gần hội để AI đem lại lợi ích cho nhân loại—nếu ta có thể làm chúng vững chắc và không thâm nhập được. Tuy chúng ta có thể dùng chính AI để làm cho các hệ thống AI vững chắc hơn, nhờ đó tăng khả năng phòng thủ trong chiến tranh mạng, nhưng AI rõ ràng cũng có thể hỗ trợ tấn công. Đảm bảo cho hàng phòng thủ luôn mạnh hơn phải là một trong những mục tiêu ngắn hạn quan trọng nhất cho sự phát triển AI—nếu không tất cả những công nghệ tuyệt vời chúng ta chế tạo có thể biến thành công cụ chống lại chúng ta!

Việc làm và lương

Tới giờ, chương này mới tập trung chủ yếu vào việc AI sẽ ảnh hưởng tới chúng ta thế nào với tư cách người tiêu dùng, khi nó cho ra đời những sản phẩm và dịch vụ mới mang tính biến đổi với giá thành vừa phải. Nhưng AI sẽ ảnh hưởng đến ta thế nào với tư cách người đi làm, khi nó biến đổi thị trường việc làm? Nếu tìm ra được cách phát triển thịnh vượng qua tự động hóa mà không khiến ai thiếu công ăn việc làm hay mục đích sống, thì ta sẽ có thể tạo ra một tương lai tuyệt vời với sự sung sướng và giàu sang chưa từng thấy cho bất cứ ai mong muốn điều đó. Hiếm có người từng suy nghĩ về điều này lâu hơn và sâu sắc hơn một trong các đồng nghiệp MIT của tôi, nhà kinh tế học Erik Brynjolfsson. Dù anh luôn ăn mặc chỉnh tề và giữ đầu tóc gọn gàng nhưng anh có dòng dõi Iceland và đôi khi tôi lại cảm thấy như anh chỉ vừa mới tĩa đi bộ râu tóc đỏ hoang dại kiểu người Viking để hòa nhập với môi trường đại học kinh doanh. Nhưng rõ ràng anh chưa từng tĩa bớt những ý tưởng “ngông” của mình; anh gọi tầm nhìn lạc quan về thị trường việc làm này là “Athens Số”. Các công dân thành Athens cổ đại có cuộc sống sung túc để có thể hưởng thụ nền dân chủ, nghệ thuật và các trò chơi. Điều này chủ yếu

là vì họ có nhiều nô lệ để làm phần lớn mọi việc. Vậy tại sao không thay thế nô lệ bằng các rô-bốt chạy bằng AI để tạo ra một xã hội số không tưởng mà mọi người đều có thể hưởng thụ? Nền kinh tế do AI lao động tạo ra theo ý Erik sẽ không chỉ loại bỏ căng thẳng và vất vả, sản xuất dư thừa mọi thứ mà ta muốn ngày nay, mà còn cung cấp nguồn sản phẩm và dịch vụ mới tuyệt vời mà người tiêu dùng ngày nay còn chưa nhận ra là họ muốn có.

Công nghệ và bất công

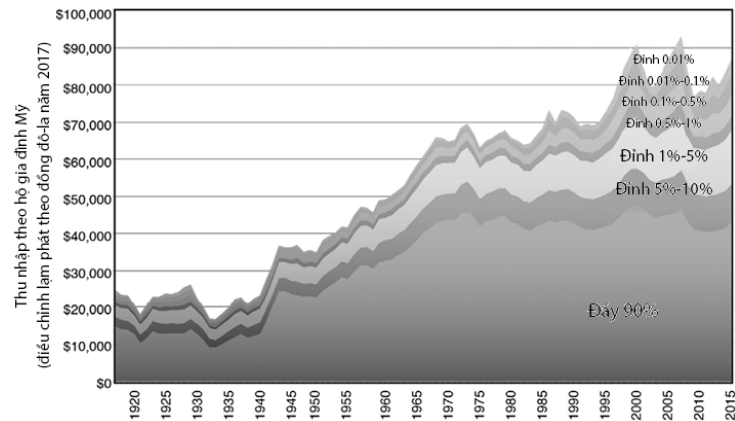
Từ thế giới hiện tại ta có thể tiến đến Athens Số của Erik nếu lương theo giờ của mọi người cứ tăng dần đều mỗi năm, để những người muốn xả hơi có thể dần dần làm việc ít đi mà vẫn tiếp tục cải thiện được mức sống. Hình 3.5 đã cho thấy đây chính là điều xảy ra ở Hoa Kỳ kể từ sau Thế chiến II đến giữa những năm 1970: mặc dù có sự bất bình đẳng thu nhập, tổng thu nhập đã trở nên lớn đến mức gần như tất cả mọi người đều có thu nhập cao hơn. Thế nhưng, cũng chính Erik đã thừa nhận rằng đã có điều thay đổi: hình 3.5 cho thấy, mặc dù nền kinh tế tiếp tục tăng trưởng và nâng cao mức thu nhập trung bình, nhưng trong 40 năm qua lợi lộc phần lớn đi vào túi những người giàu nhất, hầu hết nằm trong nhóm 1% trên cùng, còn với nhóm 90% dân số nghèo nhất thì thu nhập vẫn giậm chân tại chỗ. Sự gia tăng bất bình đẳng càng thể hiện rõ khi ta không nhìn vào thu nhập mà vào tổng giá trị tài sản. Tổng tài sản trung bình của 90% hộ gia đình nghèo nhất ở Mỹ thường nằm ở khoảng 85 nghìn đô-la Mỹ năm 2012—vẫn y như 25 năm trước đó—khi tổng tài sản sau khi điều chỉnh lạm phát của 1% giàu nhất đã tăng hơn gấp đôi trong cùng khoảng thời gian đó, lên đến 14 triệu đô-la Mỹ.⁴² Sự khác biệt trên toàn cầu lại càng khủng khiếp, khi vào năm 2013, các chuyên gia đã phân tích rằng tổng tài sản của một nửa dân số thế giới nghèo hơn (khoảng 3,6 tỉ người) lại bằng với tổng tài sản của tám người giàu nhất thế giới⁴³—một số liệu cho thấy rõ sự nghèo khổ và dễ tổn thương ở dưới đáy xã hội cũng lớn không kém gì sự giàu có ở trên đỉnh. Tại hội nghị AI ở Puerto Rico năm 2015, Erik đã phát biểu trước hội đồng đại biểu nghiên cứu AI rằng anh nghĩ tiến bộ trong AI và tự động hóa sẽ tiếp tục làm cho cái bánh kinh tế lớn hơn, nhưng sẽ

không có luật lệ kinh tế nào giúp tất cả hay phần lớn mọi người được hưởng lợi.

Mặc dù hầu hết các nhà kinh tế học nhất trí rằng bất bình đẳng thu nhập đang gia tăng, nhưng họ lại tranh cãi khá sôi nổi về nguyên nhân của xu hướng này và liệu nó có tiếp tục không. Các nhà hùng biện cánh tả trong hệ thống quan điểm chính trị thường lập luận rằng lý do chính là toàn cầu hóa và/hoặc các chính sách kinh tế như giảm thuế cho người giàu. Nhưng Erik Brynjolfsson và cộng sự Andrew McAfee ở MIT lại tranh cãi rằng nguyên nhân chính là một yếu tố khác: công nghệ.⁴⁴ Cụ thể, họ đưa ra luận điểm rằng công nghệ số đã khiến bất bình đẳng gia tăng theo ba chiều hướng khác nhau.

Thứ nhất, khi những việc làm cũ bị thay thế bằng những việc làm cần đến nhiều kỹ năng hơn, công nghệ tưởng thưởng cho những người có học vấn: kể từ giữa thập niên 1970, mức lương tăng khoảng 25% cho những người có bằng đại học, trong khi với những người chưa học xong trung học có mức lương trung bình giảm khoảng 30%.⁴⁵

Thứ hai, họ khẳng định rằng từ năm 2000, thu nhập doanh nghiệp đi vào túi những người sở hữu doanh nghiệp ngày một nhiều thay vì những người làm ở đó—và khi nào quá trình tự động hóa còn tiếp tục thì ta sẽ còn thấy những người sở hữu máy móc lấy đi miếng bánh ngày một lớn hơn. Lợi thế này của vốn so với sức lao động có thể đặc biệt quan trọng trong nền kinh tế số đang phát triển- nền kinh tế mà nhà tiên tri công nghệ Nicholas Negroponte định nghĩa là gồm các bit di chuyển, không phải nguyên tử. Ngày nay, khi mọi thứ từ sách, phim đến công cụ kê khai thuế đều đã được số hóa, ta có thể bán các bản sao của những thứ này khắp thế giới với chi phí gần như bằng không, mà không cần thuê thêm nhân viên. Điều này cho phép hầu hết lợi nhuận đi vào túi các nhà đầu tư thay vì người làm và cũng giúp giải thích lý do mặc dù tổng lợi nhuận của “Big 3” ở Thung lũng Silicon (ba công ty công nghệ lớn nhất gồm Google, Apple, Facebook) năm 2014 gần như ngang bằng tổng lợi nhuận của “Big 3” ở Detroit (GM, Ford và Chrysler) năm 1990, nhưng các ông lớn công nghệ chỉ có số lượng nhân viên bằng 1/9 nhưng lại có giá trị gấp 30 lần trên thị trường chứng khoán.⁴⁷



Hình 3.5: Hình này cho thấy nền kinh tế đã khiến mức thu nhập trung bình trong vòng một thế kỷ qua tăng trưởng như thế nào, và tỷ lệ nào trong thu nhập đang rơi vào tay những nhóm nào. Trước những năm 1970, ta có thể thấy người giàu và người nghèo cùng sánh bước cải thiện thu nhập. Sau đó, nhóm đỉnh 1% chiếm phần lớn lợi lộc còn nhóm đáy 90% hầu như không được hưởng lợi thêm.⁴⁶ Số liệu đã được điều chỉnh lạm phát theo đồng đô-la năm 2017.

Thứ ba, Erik và các cộng sự lập luận rằng nền kinh tế số thường đem lại lợi ích cho các siêu sao hơn những người khác. J. K. Rowling, tác giả bộ truyện Harry Potter đã trở thành nhà văn đầu tiên gia nhập câu lạc bộ tỷ phú. Bà giàu hơn Shakespeare rất nhiều vì các truyện của bà có thể chuyển tải dưới dạng văn bản, phim và trò chơi cho hàng tỷ người với chi phí rất thấp. Tương tự, Scott Cook đã thu về một tỷ đô-la lợi nhuận từ phần mềm kê khai thuế TurboTax. Không như các nhân viên kê khai thuế, dịch vụ này chỉ cần khách hàng tải về là có thể sử dụng được. Bởi hầu hết mọi người chỉ muốn trả rất ít tiền, hoặc chẳng muốn thanh toán, cho một phần mềm kê khai thuế hạng xoàng trên thị trường, nên thị trường sẽ chỉ có chỗ cho một số nhỏ các siêu phẩm. Điều này cũng đồng nghĩa với việc nếu tất cả các bậc cha mẹ trên thế giới khuyên con mình kế nghiệp J. K. Rowling, Gisele Bündchen, Matt Damon, Cristiano Ronaldo, Oprah Winfrey hay Elon Musk, hầu như sẽ không đứa trẻ nào thấy đây là một con đường sự nghiệp khả thi.

Hướng nghiệp cho con trẻ

Vậy ta nên hướng nghiệp cho các con thế nào? Tôi đang khuyến khích các con đi vào những nghề mà máy móc hiện nay còn đang

kém, và do đó khó có khả năng tự động hóa trong tương lai gần. Các dự báo gần đây về thời điểm việc làm sẽ bị máy móc thay thế đã vạch ra nhiều câu hỏi hay về một nghề trước khi một người quyết định đi học cho nghề đó.⁴⁸ Ví dụ:

Nghề này có yêu cầu tương tác với con người và ứng dụng tri thức xã hội không?

Nghề này có yêu cầu khả năng sáng tạo và nghĩ ra những giải pháp thông minh không?

Nghề này có yêu cầu làm việc trong môi trường khó đoán trước không?

Bạn càng trả lời “có” cho nhiều câu hỏi, lựa chọn nghề nghiệp của bạn càng có khả năng phù hợp. Điều này nghĩa là những lựa chọn tương đối an toàn sẽ bao gồm nghề giáo, y tá, bác sĩ, nha sĩ, nhà khoa học, doanh nhân, lập trình viên, kỹ sư, luật sư, nhân viên công tác xã hội, tu sĩ, nghệ sĩ, thợ làm tóc, hay thợ xoa bóp trị liệu.

Ngược lại, những nghề bao gồm các công việc mang tính lặp lại hay cấu trúc chặt chẽ trong môi trường dễ đoán trước sẽ chẳng tồn tại được lâu cho đến khi bị tự động hóa. Máy tính và người máy công nghiệp đã thay thế các việc làm đơn giản nhất như vậy từ lâu, và tiến trình công nghệ cũng đang loại bỏ thêm nhiều việc làm khác: từ tiếp thị qua điện thoại, nhân viên kho bãi, nhân viên thu ngân, vận hành đường sắt, nhân viên phụ bếp đến cả đầu bếp nướng bánh.⁴⁹ Tài xế xe tải, xe buýt, taxi và cả Uber/Lyft nhiều khả năng cũng sẽ sớm bị thay thế. Những ngành khác (bao gồm luật gia, nhân viên phân tích tín dụng, cán bộ cho vay, nhân viên kế toán và kế toán thuế) tuy không nằm trong danh sách “ngành nghề có nguy cơ tuyệt chủng”, nhưng có nhiều công đoạn được tự động hóa và do đó sẽ cần ít người hơn.

Nhưng tránh khỏi những ngành nghề được tự động hóa không phải là thách thức nghề nghiệp duy nhất. Trong kỷ nguyên số toàn cầu này, trở thành một văn sĩ, nhà làm phim, diễn viên, vận động viên, hay nhà thiết kế thời trang chuyên nghiệp là một mục tiêu mạo hiểm vì một lý do khác: mặc dù máy móc sẽ không cạnh tranh hoàn toàn với người làm những nghề này trong thời gian tới, nhưng họ sẽ

phải cạnh tranh ngày một gay gắt với những người khác trên toàn thế giới theo lý thuyết siêu sao đã nói ở trên, và sẽ chỉ có rất ít người thành công.

Trong nhiều trường hợp, đưa ra lời khuyên nghề nghiệp ở mức độ toàn ngành sẽ là điều rất hàm hồ: nhiều việc làm sẽ không bị loại bỏ hoàn toàn nhưng sẽ có nhiều công đoạn tự động hóa. Ví dụ, nếu bạn vào ngành y, đừng làm bác sĩ X quang phân tích hình ảnh để rồi bị người máy Watson của IBM thay thế, mà hãy là người bác sĩ chỉ định cho bệnh nhân đi chụp tia X, là người thảo luận kết quả với bệnh nhân, và quyết định phác đồ điều trị. Nếu bạn vào ngành tài chính, đừng làm người phân tích định lượng, chỉ biết áp dụng thuật toán vào dữ liệu để rồi bị phần mềm thay thế, mà hãy là người quản lý quỹ, sử dụng phân tích định lượng đó để đưa ra các quyết định đầu tư chiến lược. Nếu bạn vào ngành luật, đừng làm một luật gia tập sự, phải đi xem xét hàng ngàn văn bản cho bước điều tra để rồi bị tự động hóa thay thế, mà hãy là một luật sư tư vấn cho thân chủ và trình bày vụ án trước tòa.

Cho đến giờ, chúng ta đã tìm hiểu những việc mà các cá nhân có thể làm để thành công tối đa trên thị trường việc làm trong kỷ nguyên AI. Nhưng các chính phủ sẽ cần làm gì để giúp lực lượng lao động thành công? Ví dụ, hệ thống giáo dục nào sẽ trang bị tốt nhất cho một người bước vào một thị trường việc làm có các AI liên tục cải tiến nhanh chóng? Có nên tiếp tục với mô hình giáo dục hiện nay để một người đi học một đến hai thập niên, sau đó cho họ làm công việc chuyên môn trong bốn thập niên tiếp theo?⁵⁰ Hay ta nên chuyển đổi sang một mô hình cho phép mọi người đi làm trong một vài năm, đi học lại trong một năm, rồi lại đi làm tiếp trong vài năm nữa?⁵⁰ Hay giáo dục thường xuyên (có thể cung cấp trực tuyến) nên là một tiêu chuẩn của bất kỳ công việc nào?

Và chính sách kinh tế nào sẽ có ích nhất trong việc tạo ra những công việc mới tốt đẹp? Andrew McAfee lập luận rằng sẽ có nhiều chính sách hữu ích cho mục tiêu này, bao gồm đầu tư mạnh mẽ vào nghiên cứu, giáo dục và cơ sở hạ tầng, tạo điều kiện cho di cư và khuyến khích thành lập doanh nghiệp mới. Anh cảm thấy “sách Kinh

tế Đại cường nói rất rõ, nhưng không ai làm theo”, ít nhất là ở Hoa Kỳ.⁵¹

Liệu con người có đến mức không còn việc để làm?

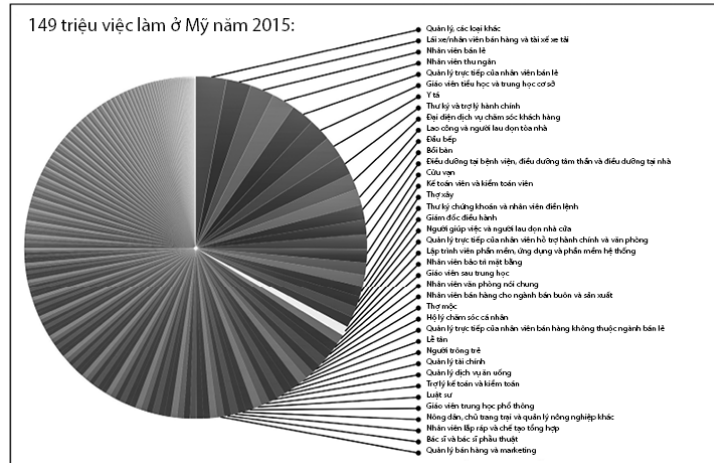
Nếu AI tiếp tục cải tiến rồi ngày càng tự động hóa nhiều công việc, điều gì sẽ xảy ra? Nhiều người lạc quan về vấn đề việc làm, tranh luận rằng các việc làm được tự động hóa sẽ được thay thế bởi những việc làm mới còn tốt hơn. Suy cho cùng, điều này cũng từng xảy ra từ thời Cách mạng Công nghiệp, khi những người phải bàn lùi bày tỏ lo ngại thất nghiệp do phát triển công nghệ.

Tuy nhiên, những người khác lại bi quan về vấn đề việc làm và tranh cãi rằng thời nay đã khác, và ngày càng nhiều người sẽ không chỉ mất việc, mà còn không thể nào tìm được việc làm nữa.⁵² Phe bi quan cho rằng, do nền kinh tế thị trường tự do đặt ra mức lương dựa trên cung cầu và lượng cung ngày càng tăng, nên lao động máy móc giá rẻ rất cuộc sẽ khiến mức lương của con người giảm thấp hơn dưới cả chi phí sinh hoạt. Vì mức lương thị trường đưa ra cho một việc làm là chi phí theo giờ trả cho bất kỳ ai hay bất kỳ cái gì thực hiện được việc đó với chi phí thấp nhất nên mức lương sẽ giảm kịch sàn bất cứ khi nào công việc đó có thể chuyển cho một nước có mức thu nhập thấp hơn hay cho một loại máy móc rẻ hơn. Trong Cách mạng Công nghiệp, ta đã tìm ra cách thay thế chân tay bằng máy móc; mọi người dần chuyển sang làm những việc trả lương cao và yêu cầu trí óc hơn. Các công việc trí óc dần thay thế các công việc chân tay. Ngày nay, ta lại đang tìm cách thay thế trí óc bằng máy móc. Nếu quá trình này thành công, thì còn lại gì để chúng ta làm?

Một số người phe lạc quan cho rằng sau khi công việc chân tay và trí óc bị thay thế thì nhu cầu công việc sáng tạo sẽ bùng nổ, nhưng phe bi quan lại phản bác rằng sáng tạo cũng chỉ là một quá trình trí óc, vậy nên cuối cùng AI cũng sẽ làm chủ. Những người khác ở phe lạc quan lại hy vọng rằng trong tương lai sẽ bùng nổ những nghề nghiệp do công nghệ sinh ra mà ta còn chưa nghĩ tới. Suy cho cùng, những người sống ở thời Cách mạng Công nghiệp cũng đâu thể tưởng tượng sau này con cháu mình làm nghề thiết kế web hay lái xe Uber? Nhưng phe bi quan lại phản pháo rằng, đây chỉ là suy nghĩ viễn vông chưa có nhiều dữ liệu thực nghiệm hậu thuẫn. Họ chỉ ra

chúng ta cũng đã có thể lập luận như vậy từ một thế kỷ trước, trước khi có cách mạng máy tính; rằng chúng ta đã dự đoán hầu hết nghề nghiệp ngày nay sẽ mới mẻ chưa từng có và những nghề nghiệp chưa ai tưởng tượng được này đều do công nghệ sinh ra. Dự đoán này đã thất bại thảm hại, như minh họa trong hình 3.6: hầu hết các ngành nghề ngày nay đã tồn tại từ cả thế kỷ trước, và khi phân loại chúng theo lượng việc làm tạo ra, ta cũng phải lần xuống tới mục thứ 21 trong danh sách mới thấy một nghề mới: lập trình phần mềm, chỉ chiếm tỷ trọng dưới 1% thị trường việc làm ở Mỹ.

Ta có thể hiểu rõ hoàn cảnh này hơn bằng cách xem lại khung cảnh trí tuệ con người ở hình 2.2, chương 2, với chỗ cao tượng trưng cho mức độ khó với máy móc khi thực hiện nhiều nhiệm vụ khác nhau và mực nước biển đang dâng lên cho thấy những gì máy móc hiện nay có thể làm được. Trên thị trường việc làm hiện tại, xu hướng chính của chúng ta không phải là dịch chuyển sang những nghề nghiệp hoàn toàn mới. Thay vào đó, chúng ta đang tụ tập vào những khu vực địa hình chưa bị cơn thủy triều công nghệ san lấp trong hình 2.2! Hình 3.6 lại cho thấy việc này không chỉ tạo ra một hòn đảo duy nhất mà là cả một quần đảo phức tạp, trong đó những đảo nhỏ và đảo san hô vòng tương ứng với những nghề giá trị cao mà máy móc vẫn chưa thể làm được với giá rẻ như con người. Đó không chỉ là những nghề nghiệp công nghệ cao như lập trình phần mềm mà còn là một loạt những nghề công nghệ thấp sử dụng đến những kỹ năng xã hội và sự khéo léo vượt trội của chúng ta, từ xoa bóp trị liệu đến diễn xuất. Liệu AI có thể vượt qua con người ở những nhiệm vụ trí óc ấy nhanh chóng đến mức ta chỉ còn có thể làm những ngành nghề công nghệ thấp ấy? Gần đây, một người bạn đã đùa với tôi rằng có lẽ nghề nghiệp cuối cùng tồn tại trên đời sẽ là nghề nghiệp nguyên thủy: bán dâm. Nhưng khi anh ấy nói điều đó với một nhà rô-bốt học người Nhật, người đó đã phản bác: “Không, rô-bốt giỏi chuyện ấy lắm đấy!”



Hình 3.6: Biểu đồ tròn cho thấy các ngành nghề mà 149 triệu người Mỹ làm năm 2015, với 535 loại việc làm được Cục Thống kê Lao động Hoa Kỳ sắp xếp theo độ phổ biến.⁵³ Tất cả ngành nghề với hơn một triệu người lao động đều được ghi lại tại đây. Công nghệ máy tính chỉ tạo ra các ngành nghề mới ở các thứ hạng sau vị trí 21. Hình minh họa này có liên quan tới bài phân tích của Federico Pistono.⁵⁴

Những người bi quan về việc làm khẳng định điểm kết thúc đã rõ ràng: toàn bộ quần đảo sẽ chìm trong biển nước, và con người sẽ không thể thực hiện công việc nào với giá thành rẻ hơn máy móc nữa. Trong cuốn sách năm 2007 *Farewell to Alms* (Giã từ bố thí³⁷), nhà kinh tế học người Mỹ gốc Scotland Gregory Clark chỉ ra rằng ta có thể nhìn thấu tương lai việc làm đôi chút nếu so sánh tình cảnh của bản thân với loài ngựa. Thử tưởng tượng cảnh hai con ngựa nhìn về phía một chiếc xe ô tô đời đầu năm 1900 và suy nghĩ về tương lai của chúng xem:

“Mình lo sau này sẽ thất nghiệp vì công nghệ mất.”

“Hí, hí, đừng có bàn lùi: tổ tiên ta từng nói y như thế khi đầu máy hơi nước lấy mất việc làm của chúng ta trong công nghiệp và tàu lửa lấy đi việc kéo xe ngựa chở khách. Nhưng giờ chúng ta vẫn có nhiều việc làm hơn bao giờ hết, và còn tốt hơn nữa đây thôi: mình thà kéo xe chở hàng nhẹ nhàng qua thị trấn, hơn là dành cả ngày đi vòng quanh để chạy một cái bơm khơi mỏ ngu ngốc.”

“Nhưng nhờ cái thứ động cơ đốt trong này thật sự sẽ làm nên chuyện thì sao?”

“Mình chắc chắn sẽ có nhiều nhiều việc mới cho loài ngựa mà ta còn chưa hình dung ra. Trước nay bao giờ chẳng vậy, giống hời người ta phát minh ra bánh xe và cái cày ấy.”

Nhưng buồn thay, những việc chẳng-thể-hình-dung cho loài ngựa kia không hề xuất hiện. Những con ngựa chẳng-ai-cần đều bị giết, chứ không ai bổ nhiệm chúng vào công dụng khác. Điều này khiến cho số lượng ngựa ở Mỹ tụt từ 26 triệu con vào năm 1915 xuống còn khoảng 3 triệu vào năm 1960.⁵⁵ Liệu những khối óc cơ khí, như những cơ bắp cơ khí từng loại bỏ vai trò loài ngựa có khiến loài người rơi vào số phận tương tự?

Không cần việc làm vẫn có thu nhập

Vậy thì ai đúng, những người nói rằng những việc làm bị tự động hóa sẽ được thay thế bằng những việc làm tốt hơn, hay những người nói phần lớn nhân loại rồi sẽ không thể kiếm được việc làm nữa? Nếu AI vẫn tiến bộ mà không suy giảm, thì cả hai phe đều có thể đúng: một phe đúng ở tầm ngắn hạn còn phe kia đúng ở tầm dài hạn. Nhưng dù người ta thường bàn về tương lai mất việc làm với sắc thái u ám, tương lai này cũng không hẳn xấu! Những người bàn lùi từng ám ảnh về một số việc làm nhất định mà bỏ qua khả năng rằng những việc làm khác cũng đem lại giá trị xã hội tương đương. Tương tự như vậy, những người đang ám ảnh về việc làm ngày nay dường như có quan điểm hơi hạn hẹp: chúng ta muốn có việc làm vì việc làm đem lại cho ta thu nhập và mục đích sống, nhưng với lượng tài nguyên phong phú mà máy móc sản xuất ra, chúng ta sẽ có thể tìm ra những cách khác để thỏa mãn cả thu nhập và mục đích sống mà không cần có việc làm. Điều tương tự cũng xảy ra nếu với hoàn cảnh của loài ngựa; hệ quả đó đã không khiến cho loài ngựa tuyệt chủng. Thay vào đó, kể từ năm 1960, số lượng ngựa lại tăng lên hơn ba lần nhờ một hệ thống “phúc lợi xã hội” cho ngựa: mặc dù chúng không thể tự trả chi phí nuôi mình, người ta vẫn quyết định chăm sóc cho ngựa, nuôi chúng để bầu bạn cho vui hay thi đấu thể thao. Liệu ta có thể làm việc tương tự để chăm sóc cho những đồng loại con người khốn khó?

Hãy bắt đầu với vấn đề về thu nhập: chỉ cần chia lại một phần nhỏ của chiếc bánh kinh tế đang tăng trưởng cũng đủ để giúp tất cả mọi người khá giả hơn. Nhiều người cho rằng, chúng ta không những có thể mà còn nên làm thế. Trong buổi đối thoại năm 2016, khi Moshe Vardi phát biểu về thôi thúc đạo đức phải giúp con người bằng công nghệ AI, tôi đã tranh luận rằng cũng nên cất tiếng gọi lương tri để tuyên truyền ủng hộ cho việc dùng AI sao cho có lợi, trong đó bao gồm chia sẻ khối tài sản. Cùng tham gia cuộc đối thoại này, Erik Brynjolfsson cho rằng, “nếu khối tài sản mới sản sinh này không thể giúp chúng ta ngăn một nửa dân số thế giới nghèo đi thì thật đáng xấu hổ!”

Có rất nhiều đề xuất chia sẻ tài sản khác nhau, mỗi đề xuất đều có những người ủng hộ và nhạo báng. Đơn giản nhất là thu nhập cơ bản, theo đó mỗi người sẽ được nhận một khoản tiền hàng tháng vô điều kiện. Một số nước như Canada, Phần Lan và Hà Lan đang lên kế hoạch hoặc đang thí điểm hình thức này. Những người ủng hộ lập luận rằng thu nhập cơ bản sẽ hiệu quả hơn các biện pháp như chi trả phúc lợi cho người có hoàn cảnh khó khăn, vì nó loại trừ các thủ tục quan liêu phiền phức để xác định ai đủ tiêu chuẩn nhận phúc lợi. Phúc lợi dựa trên hoàn cảnh và nhu cầu cũng đã bị chỉ trích vì làm mất động cơ làm việc, nhưng tất nhiên điều này không còn ý nghĩa trong một tương lai không có việc làm và không ai làm việc.

Chính phủ có thể giúp đỡ người dân không chỉ bằng cách cho họ tiền, mà còn bằng cách trợ giá hoặc miễn phí các dịch vụ như cầu đường, công viên, giao thông công cộng, trông trẻ, giáo dục, y tế, dưỡng lão và kết nối Internet; quả thực, nhiều chính phủ hiện đã cung cấp hầu hết những dịch vụ này. Trái ngược với thu nhập cơ bản, các dịch vụ do chính phủ đầu tư đã hoàn thành được hai mục tiêu riêng rẽ: làm giảm chi phí sinh hoạt của người dân tạo ra việc làm. Ngay cả trong một tương lai mà máy móc làm mọi việc tốt hơn con người, chính phủ vẫn có thể lựa chọn trả tiền cho mọi người để làm trong các lĩnh vực như chăm sóc trẻ em và người già, v.v. thay vì thuê rô-bốt.

Điều thú vị là tiến bộ công nghệ rồi sẽ có thể cung cấp miễn phí nhiều sản phẩm và dịch vụ có giá trị mà không cần chính phủ can thiệp. Ví dụ, trước đây người ta phải chi tiền để mua bách khoa toàn

thư địa đồ, gửi thư và gọi điện thoại, nhưng giờ bất kỳ ai có kết nối internet cũng có thể sử dụng miễn phí những thứ này— cùng với dịch vụ hội thảo truyền hình trực tuyến, chia sẻ hình ảnh, truyền thông xã hội, các khóa học trực tuyến và vô số các dịch vụ mới khác. Nhiều thứ có thể rất quý giá đối với một số người, như một liều kháng sinh lúc nguy cấp, cũng đã trở nên cực kỳ phải chăng. Vậy là nhờ công nghệ, nhiều người nghèo ngày nay cũng có thể tiếp cận những thứ mà người giàu nhất thế giới ngày xưa hẳn còn thiếu. Theo một số người, điều này cho thấy thu nhập cần cho một cuộc sống tương đối đang giảm dần.

Nếu một ngày nào đó máy móc có thể sản xuất tất cả hàng hóa và dịch vụ hiện nay với giá thành tối thiểu, thì rõ ràng tài sản tạo thành sẽ khiến tất cả mọi người khá giả hơn. Nói cách khác, ngay cả với mức thuế thấp, các chính phủ cũng vẫn có thể chi trả cho thu nhập cơ bản và dịch vụ miễn phí. Nhưng việc tài sản có thể chia sẻ tất nhiên không đồng nghĩa rằng sẽ có người chia sẻ, thậm chí vẫn có những bất đồng chính trị mạnh mẽ về việc liệu chúng ta có nên chia sẻ hay không. Như ta thấy ở trên, nước Mỹ có vẻ đang đi ngược xu hướng: ở đây đang có những nhóm người khánh kiệt dần qua từng thập kỷ. Các chính sách nhằm tìm cách chia sẻ khối tài sản ngày càng tăng của xã hội sẽ ảnh hưởng đến tất cả mọi người, nên nội dung bàn luận về phương pháp xây dựng nền kinh tế trong tương lai cần bao gồm mọi tầng lớp người, chứ không nên giới hạn ở những người nghiên cứu AI, rô-bốt học và kinh tế học.

Nhiều người cũng đã tranh luận rằng giảm bất bình đẳng thu nhập là một ý tưởng hay ngay vào lúc này chứ không chỉ trong tương lai có AI ngự trị. Mặc dù lập luận chính của họ thiên về mặt đạo đức, nhưng cũng có bằng chứng cho thấy càng bình đẳng thì nền dân chủ càng hoạt động tốt: khi phát sinh một tầng lớp lớn những người trung lưu có học thức, các cử tri sẽ khó bị dắt mũi hơn, và các cá nhân hay các doanh nghiệp sẽ khó dùng tiền gây ảnh hưởng phi pháp lên chính phủ hơn. Một nền dân chủ tốt hơn cũng giúp nền kinh tế được quản lý tốt hơn, ít tham nhũng hơn, hiệu quả hơn và tăng trưởng nhanh hơn, cuối cùng sẽ đem lại lợi ích cho mọi người.

Không cần việc làm vẫn có mục đích

Việc làm có thể đem lại nhiều giá trị hơn chỉ tiền bạc. Voltaire đã viết năm 1759 là “công việc làm đầy lùi ba thói hư tật xấu: buồn chán, trụy lạc và túng thiếu.” Ngược lại, đem lại thu nhập cho mọi người vẫn chưa đủ để đảm bảo họ sống tốt. Các hoàng đế La Mã đã cung cấp cả bánh mì và hí trường để khiến dân chúng hài lòng, còn Chúa Jesus thì nhấn mạnh những nhu cầu phi vật chất trong Kinh thánh “Người ta sống chẳng phải chỉ nhờ bánh mà thôi”. Vậy chính xác thì việc làm đem lại những điều giá trị nào ngoài tiền bạc, và có cách nào khác để một xã hội không việc làm cũng đem lại những điều đó?

Câu trả lời rõ ràng sẽ phức tạp, vì có người ghét công việc của mình, có người yêu thích. Hơn nữa, nhiều trẻ em, học sinh và người nội trợ vẫn sống tốt mà không cần việc làm, trong khi lịch sử chứa đầy những câu chuyện về các cậu ấm, cô chiêu ăn chơi bạt mạng, để rồi phải chịu đựng sự chán chường và trống trải. Một phân tích tổng hợp năm 2012 cho thấy thất nghiệp thường có hậu quả tiêu cực lâu dài lên sức khỏe và hạnh phúc, trong khi nghỉ hưu thì đem lại cả tác động tiêu cực lẫn tích cực.⁵⁶ Ngành tâm lý học tích cực đang phát triển đã xác định một số yếu tố thúc đẩy cảm giác khỏe mạnh, hạnh phúc và có mục đích sống, và thấy rằng một số (nhưng không phải tất cả!) công việc có thể đem lại nhiều yếu tố như vậy, bao gồm:⁵⁷

Một mạng lưới xã hội gồm bạn bè và đồng nghiệp
Một lối sống lành mạnh và đạo đức
Lòng tự trọng, tự tôn, tự tin và cảm giác “trôi chảy” khi làm một thứ mà mình giỏi dẫn đến sự hài lòng
Cảm giác được cần đến và có khả năng tạo nên khác biệt
Cảm giác có ý nghĩa từ việc là một phần và phục vụ cho một mục đích vĩ đại hơn bản thân

Những điều này cho chúng ta lý do để lạc quan, vì chúng ta có thể nhận tất cả những thứ kia ngoài môi trường công sở, chẳng hạn qua các hoạt động thể thao, các thú vui sở thích và học hành, cùng với gia đình, bạn bè, đồng đội, các câu lạc bộ, nhóm hoạt động cộng đồng, trường học, tổ chức tôn giáo và nhân văn, phong trào chính trị và các cơ sở khác. Do đó, để tạo ra một xã hội ít việc làm mà vẫn thịnh vượng chứ không thoái hóa tự hủy hoại, chúng ta cần hiểu

cách duy trì và phát triển những hoạt động đem lại sức khỏe và hạnh phúc đó. Để hiểu được những điều này, ta không chỉ cần các nhà khoa học và kinh tế học hợp tác, mà còn cần các nhà tâm lý học, xã hội học và các nhà giáo. Nếu ta dùng một phần tài sản mà AI sinh ra để nghiêm túc nỗ lực tạo ra sức khỏe và hạnh phúc cho tất cả mọi người thì xã hội sẽ có thể hưng thịnh như chưa từng thấy. Ở mức độ tối thiểu, xã hội đó khiến mọi người hạnh phúc như thể đã trao cho họ công việc mà họ hằng mơ ước, nhưng một khi mọi người không còn bị buộc phải tạo ra thu nhập từ công việc cá nhân, thì hẳn sẽ không có giới hạn nào đối với những gì mà xã hội có thể đem lại.

Trí tuệ trình độ con người?

Trong chương này, ta đã tìm hiểu về cách AI có thể cải thiện lớn lao cuộc sống của chúng ta trong tương lai gần, miễn là chúng ta có kế hoạch trước và tránh được nhiều cạm bẫy khác nhau. Nhưng còn về lâu dài thì sao? Tiến bộ trong phát triển AI liệu có chững lại do những chướng ngại không thể vượt qua, hay rồi các nhà nghiên cứu AI sẽ thành công trong mục tiêu xây dựng trí tuệ nhân tạo tổng quát trình độ con người? Chúng ta đã thấy trong chương trước cách các định luật vật lý cho phép những khối vật chất phù hợp có thể ghi nhớ, tính toán và học tập ra sao, cũng như việc những khối vật chất này không bị các định luật vật lý giới hạn, để rồi một ngày nào đó trở nên thông minh hơn cả khối vật chất trong đầu chúng ta. Nhưng chúng ta chưa thể biết rõ khi nào, hay có thể nào, con người thành công trong việc chế tạo ra một AGI siêu phàm như vậy. Từ chương đầu tiên, chúng ta đều thấy rằng bản thân chưa biết, vì ngay cả những chuyên gia AI hàng đầu thế giới cũng chưa thể thống nhất. Hầu hết, đưa ra những đánh giá từ vài thập kỷ đến vài thế kỷ; có người còn dự đoán là chẳng bao giờ. Dự đoán tương lai thật khó, vì khi bạn đang khám phá vùng đất chưa ai đặt chân đến, bạn sẽ chẳng biết có bao nhiêu ngọn núi đang ngăn cách bạn với đích đến. Thông thường bạn sẽ chỉ thấy đỉnh núi gần nhất, và phải trèo qua nó thì mới phát hiện ra chướng ngại vật tiếp theo.

Vậy sớm nhất là bao giờ? Giả sử quả thực ta có biết cách thiết kế AGI trình độ con người tốt nhất bằng phần cứng máy tính hiện nay thì lượng phần cứng hiện tại cũng cần đủ khả năng cung cấp năng lực tính toán thô sơ cần thiết. Vậy năng lực tính toán của bộ não con

người bằng bao nhiêu khi đo bằng đơn vị bit và FLOPS như đã đề cập trong chương 2?³⁸ Đây là một câu hỏi khó thú vị, và câu trả lời phụ thuộc phần nhiều vào cách ta đặt câu hỏi:

Câu hỏi 1: Chúng ta cần bao nhiêu FLOPS để mô phỏng một bộ não?

Câu hỏi 2: Chúng ta cần bao nhiêu FLOPS để đạt trí tuệ mức con người?

Câu hỏi 3: Bộ não con người có thể thực hiện bao nhiêu FLOPS?

Đã có rất nhiều bài nghiên cứu về câu hỏi 1 và đáp án thường ở quanh mức 100 petaFLOPS, tức 10^{17} FLOPS.⁵⁸ Con số này tương đương với năng lực tính toán của máy Sunway TaihuLight (hình 3.7), siêu máy tính nhanh nhất thế giới năm 2016 với tổng chi phí khoảng 300 triệu đô-la Mỹ. Kể cả khi có thể dùng nó để mô phỏng bộ não của một nhân viên lành nghề, thì ta cũng chỉ có thể thu lợi nhuận nếu máy TaihuLight thực hiện được công việc của người đó với giá thấp hơn lương theo giờ của họ. Ta có thể phải chi trả thậm chí nhiều hơn, vì nhiều nhà khoa học tin rằng để mô phỏng chính xác trí tuệ của bộ não thì ta không thể coi nó là một mô hình mạng nơ-ron toán học đơn giản như trong chương 2. Thay vào đó, có vẻ ta cần mô phỏng nó ở cấp độ từng phân tử hay thậm chí ở mức hạ nguyên tử³⁹, và việc đó sẽ cần nhiều FLOPS hơn hẳn.

Câu trả lời cho câu hỏi 3 lại dễ hơn: Tôi cực kỳ kém trong việc nhân các số có 19 chữ số, kể cả có làm bằng giấy và bút. Nghĩa là, tính ra bộ não tôi chỉ tính toán được ở mức 0,01 FLOPS – chỉ bằng 10-19 lần đáp án của câu hỏi 1! Sự khác biệt khổng lồ này là do não bộ và siêu máy tính được tối ưu hóa cho những nhiệm vụ cực kỳ khác nhau. Khác biệt tương tự cũng nảy sinh giữa các câu hỏi sau:

Máy kéo có thể thực hiện công việc của xe đua Công thức Một đến mức nào?

Xe đua Công thức Một có thể thực hiện công việc của máy kéo đến mức nào?

Vậy, trong hai câu hỏi về FLOPS mà ta đang tìm cách trả lời, câu nào sẽ dự đoán được tương lai của AI? Chẳng câu nào cả! Nếu ta muốn mô phỏng não người, ta sẽ quan tâm đến câu hỏi đầu tiên, nhưng để chế tạo AGI trình độ con người thì điều quan trọng lại là câu hỏi thứ hai. Chúng ta vẫn chưa có đáp án cho câu hỏi này, nhưng các phương án như sửa đổi phần mềm cho phù hợp với máy tính ngày nay hoặc chế tạo phần cứng giống não bộ hơn (hiện việc chế tạo những chip dạng hệ thần kinh đặc biệt đang tiến triển rất nhanh có thể rẻ hơn mô phỏng não bộ rất nhiều).

Hans Moravec đã ước đoán câu trả lời bằng một phép so sánh khách quan nhất, sử dụng một tính toán mà cả não bộ và máy tính ngày nay đều có thể thực hiện hiệu quả: Các phép xử lý hình ảnh cấp thấp mà võng mạc con người thực hiện ở đáy mắt trước khi gửi kết quả đến não qua dây thần kinh thị giác.⁵⁹ Ông nhận thấy, để mô phỏng các tính toán của võng mạc trên một máy tính thông thường cần đến khoảng một tỉ FLOPS và toàn bộ não bộ có khả năng thực hiện lượng tính toán gấp khoảng 10.000 lần so với võng mạc (dựa trên so sánh thể tích và số lượng nơ-ron thần kinh), từ đó ta suy ra năng lực tính toán của bộ não vào khoảng 10^{13} FLOPS – tương đương năng lực của một máy tính với cấu hình tối ưu trị giá 1000 đô-la Mỹ năm 2015!



Hình 3.7: Sunway TaihuLight, siêu máy tính nhanh nhất thế giới năm 2016, với năng lực tính toán thô có thể cho là vượt qua con người.

Tóm lại, chúng ta không thể đảm bảo tuyệt đối rằng có thể chế tạo AGI trình độ con người trong thời đại này – hay liệu có khả năng chế tạo ra nó. Nhưng ta cũng không thể lập luận như đinh đóng cột rằng những điều nêu trên là bất khả thi. Chúng ta không thể viện đến lý do là thiếu năng lực phần cứng hay quá đắt đỏ được nữa. Tuy không biết vạch đích còn cách bao xa nếu xét về mặt thiết kế, thuật toán và

phần mềm, nhưng hiện tại chúng ta vẫn đang tiến bộ nhanh chóng và các thách thức tồn tại vẫn được một cộng đồng các nhà nghiên cứu AI toàn cầu đang ngày càng lớn mạnh giải quyết. Nói cách khác, ta không thể loại bỏ hoàn toàn khả năng AGI không chỉ đạt mà còn vượt qua trình độ con người. Do đó chúng ta sẽ dành chương tiếp theo để tìm hiểu khả năng này cũng như con đường nó dẫn đến!

TÓM TẮT

- *Tiến bộ AI trong tương lai gần có tiềm năng cải thiện đáng kể cuộc sống của chúng ta trên nhiều phương diện, từ việc giúp cuộc sống cá nhân, điện lưới và thị trường tài chính vận hành hiệu quả hơn tới các khả năng cứu người bằng xe tự lái, rô-bốt phẫu thuật và hệ thống chẩn đoán AI.*

- *Khi chúng ta cho phép AI kiểm soát các hệ thống trong thế giới thực, học cách làm cho AI ổn định hơn để thực hiện các mục đích của chúng ta là điều tối quan trọng. Để làm điều này, ta cần giải quyết các vấn đề kỹ thuật khó khăn liên quan đến kiểm định, thẩm định, bảo an và kiểm soát.*

- *Nhu cầu cải thiện tính ổn định này đặc biệt cấp thiết đối với hệ thống vũ khí do AI điều khiển, vốn có nguy cơ rất cao.*

- *Nhiều nhà nghiên cứu AI và rô-bốt hàng đầu đã kêu gọi lập nên một hiệp ước quốc tế nhằm cấm một vài loại vũ khí tự hành cụ thể, để tránh một cuộc chạy đua vũ trang vượt tầm kiểm soát mà cho phép bất cứ ai có khả năng tài chính và động cơ dễ dàng tiếp cận các cỗ máy ám sát.*

- *AI có thể giúp hệ thống pháp lý công bằng và hiệu quả hơn nếu chúng ta có thể tìm ra cách làm cho rô-bốt-thẩm-phán vận hành minh bạch và không thiên vị.*

- *Luật pháp cần nhanh chóng cập nhật để theo kịp tiến bộ AI; việc này đặt ra các câu hỏi khó khăn về mặt pháp lý liên quan đến khía cạnh quyền riêng tư, trách nhiệm pháp lý và các quy tắc cho AI.*

- *Trước khi chúng ta cần lo việc những cỗ máy thông minh thay thế chúng ta hoàn toàn, chúng có lẽ sẽ dần thay thế chúng ta trên thị trường việc làm.*

- *Đây không nhất thiết là xu hướng tiêu cực, chỉ cần xã hội phân phối lại một phần khối tài sản mà AI tạo ra để giúp mọi người sống*

tốt hơn.

- Nhiều nhà kinh tế cho rằng, nếu không có điều trên thì sự bất bình đẳng sẽ tăng lên đáng kể.

- Nếu quy hoạch trước, một xã hội có tỉ lệ việc làm thấp, nơi con người tìm được mục đích từ các hoạt động khác thay vì việc làm, sẽ phát triển không chỉ về mặt kinh tế.

- Lời khuyên về nghề nghiệp cho giới trẻ ngày nay: Hãy chọn các nghề mà máy móc không giỏi — như những việc liên quan đến con người, sự bất định và tính sáng tạo.

- Có một khả năng không dễ bỏ qua là AGI sẽ tiến bộ tiệm cận trình độ con người và thậm chí vượt qua – chúng ta sẽ khám phá điều này trong chương tiếp theo!

Chương 4

Bùng nổ trí tuệ?

Nếu một cỗ máy có khả năng tư duy, có lẽ nó sẽ tư duy thông minh hơn chúng ta; khi ấy chúng ta sẽ ra sao? Kể cả khi chúng ta có thể giữ máy móc ở vị thế phục tùng ... chúng ta nên, với tư cách giống loài, khiêm nhường thôi.

Alan Turing, 1951

Cỗ máy siêu thông minh đầu tiên sẽ là phát minh cuối cùng loài người cần thực hiện, với điều kiện cỗ máy đó đủ nghe lời để cho chúng ta biết cách kiểm soát nó.

Irving J. Good, 1965

Bởi chúng ta không thể loại bỏ hoàn toàn khả năng sẽ xây dựng nên một AGI trình độ con người, hãy dành trọn chương này khám phá con đường mà khả năng này dẫn tới. Chúng ta có thể bắt đầu với vấn đề nan giải mà ít ai muốn đề cập đến sau đây:

Liệu AI có thể thực sự tiếp quản thế giới, hoặc giúp con người làm điều đó?

Nếu bạn thường đảo mắt mỗi khi mọi người nói tới khả năng rô-bốt bùng nổ như trong phim Kẻ huỷ diệt (Terminator) thống trị, thì bạn hoàn toàn có lý: đây là một viễn cảnh thực sự thiếu thực tế và ngớ ngẩn. Những con rô-bốt mà Hollywood vẽ ra này không thông minh hơn chúng ta nhiều, và thậm chí còn chẳng thành công. Theo ý kiến của tôi, mối nguy từ kịch bản Kẻ huỷ diệt không phải ở tính khả thi của nó, mà ở chỗ nó sẽ lệch khỏi những rủi ro và cơ hội thực đến từ AI. Để chuyển tiếp từ cuộc sống ngày nay đến thời kỳ AGI hỗ trợ tiếp quản thế giới, chúng ta cần thực hiện ba bước hợp lý:

- Bước 1: Xây dựng AGI trình độ con người.
- Bước 2: Sử dụng AGI này để tạo ra siêu trí tuệ.
- Bước 3: Sử dụng hoặc giải phóng siêu trí tuệ này để tiếp quản thế giới.

Trong chương trước, chúng ta đã thấy rất khó để cho rằng bước 1 vĩnh viễn không bao giờ xảy ra. Chúng ta cũng thấy rằng nếu hoàn

thành bước 1 thì cũng sẽ khó mà cho là bước 2 vô vọng, bởi AGI thành phẩm kia sẽ đủ khả năng để thiết kế đệ quy ra các AGI ngày càng tốt hơn và chỉ bị giới hạn bởi các định luật vật lý – thứ vẫn cho phép tạo ra trí tuệ vượt xa trình độ con người. Cuối cùng, vì loài người chúng ta đã thống trị các dạng sống khác trên Trái đất nhờ trí thông minh vượt bậc, nên cũng sẽ có khả năng siêu trí tuệ sẽ vượt mặt mà thống trị chúng ta như vậy.

Tuy nhiên, những lập luận có vẻ thuyết phục này lại chung chung và thiếu cụ thể đến khó chịu, trong khi điều quan trọng thường ẩn nấp trong các chi tiết. Vậy liệu AI có thực sự dẫn tới thống trị thế giới? Để khám phá câu hỏi này, hãy quên đi những Kẻ huỷ diệt ngớ ngẩn; thay vào đó, bạn có thể xem xét một vài kịch bản chi tiết về những việc có khả năng xảy ra. Sau đó, chúng ta sẽ mổ xẻ và công kích những bối cảnh này, vì vậy hãy đọc chúng bằng đôi mắt hoài nghi – chúng sẽ chủ yếu thể hiện việc chúng ta chẳng biết gì nhiều về điều sẽ và sẽ không xảy ra, cũng như cho thấy rằng phạm vi các khả năng xảy ra là cực lớn. Viễn cảnh đầu tiên sẽ nằm ở phạm vi khả năng nhanh chóng và ấn tượng nhất. Theo tôi, các viễn cảnh này sẽ rất đáng để đi sâu khám phá – không chỉ vì chúng khả thi nhất, mà còn bởi nếu không thể thuyết phục bản thân rằng chúng cực khó xảy ra thì chúng ta cần hiểu những khả năng này đủ rõ để đề phòng trước khi quá muộn, và ngăn chặn chúng dẫn đến hậu quả xấu.

Phần mở đầu của cuốn sách này là một viễn cảnh cho thấy con người sử dụng siêu trí tuệ để chiếm lĩnh thế giới. Nếu bạn chưa đọc, vui lòng quay lại và đọc nó bây giờ. Ngay cả khi bạn đã đọc rồi, hãy cân nhắc xem lướt qua nó lần nữa luôn, để nó rõ ràng hơn trong đầu ta trước khi ta phê bình và sửa đổi nó.

* * *

Chúng ta sẽ sớm khám phá các lỗ hổng nghiêm trọng trong kế hoạch của hội Omega, nhưng giả sử trong giây lát rằng kế hoạch này thành công thì bạn sẽ thấy thế nào? Bạn có muốn chứng kiến hoặc ngăn chặn việc này không? Đây hẳn sẽ là chủ đề tuyệt vời để trò chuyện sau bữa tối! Điều gì sẽ xảy ra sau khi hội Omega củng cố vững chắc quyền kiểm soát thế giới của họ? Điều đó phụ thuộc vào

mục tiêu của họ, vốn là điều mà tôi thành thực không biết. Nếu được quyết định, bạn sẽ tạo ra tương lai như thế nào? Chúng ta sẽ khám phá một loạt các lựa chọn trong chương 5.

Chế độ chuyên chế

Giờ, ta giả sử ngài CEO điều khiển hội Omega có mục tiêu lâu dài tương tự như của Adolf Hitler hay Joseph Stalin. Chúng ta chỉ biết rằng điều này hoàn toàn có thể xảy ra và ông ta chỉ đang giữ bí mật về mục tiêu này cho tới khi có đủ quyền lực để thực thi. Kể cả khi mục tiêu ban đầu của vị CEO có cao quý, thì như Nam tước Acton đã cảnh báo vào năm 1997, “quyền lực có xu hướng tha hóa và quyền lực tuyệt đối thì sẽ tha hóa tuyệt đối.” Ví dụ, ông ta có thể dễ dàng sử dụng Prometheus để tạo ra nhà nước giám sát hoàn hảo. Nếu như chương trình do thám chính phủ bị Edward Snowden tiết lộ từng tiến đến mức “toàn diện” – tức đạt khả năng ghi lại tất cả các thông tin liên lạc điện tử để phân tích sau nếu cần – thì Prometheus có thể nâng cao tới mức có thể hiểu rõ mọi liên lạc điện tử. Khi đọc tất cả các email và văn bản từng được gửi, nghe mọi cuộc điện thoại, xem tất cả các băng ghi hình giám sát và camera giao thông, phân tích mọi giao dịch thẻ tín dụng và nghiên cứu mọi hành vi trực tuyến, Prometheus sẽ có hiểu biết sâu sắc về suy nghĩ và hành động của mọi con người trên Trái đất. Nhờ phân tích dữ liệu tháp di động, nó sẽ biết vị trí của phần lớn mọi người ở mọi thời điểm. Tất cả những điều này đều nằm trên giả định rằng nó sử dụng công nghệ thu thập dữ liệu hiện tại, nhưng Prometheus có thể dễ dàng phát minh ra các món đồ công nghệ cũng như các thiết bị đeo trên người được ưa chuộng và có khả năng loại bỏ quyền riêng tư của người dùng; chúng có thể ghi lại và tải lên mọi thứ họ nghe và nhìn thấy cũng như phản ứng của họ.

Với công nghệ siêu phàm, bước chuyển từ một quốc gia chuyên chế hoàn hảo tới một quốc gia chuyên chế cảnh sát hoàn hảo sẽ rất ngắn ngủi. Ví dụ, chỉ cần viện lý do chống tội phạm và khủng bố hay để đến chỗ có những người cần được cấp cứu, các chính phủ có thể yêu cầu tất cả mọi người đeo “vòng tay an ninh”, như đồng hồ Apple Watch kèm chức năng báo cáo liên tục về vị trí, tình trạng sức khỏe

và nội dung các cuộc trò chuyện. Hành động tháo bỏ hoặc vô hiệu hóa trái phép sẽ khiến chiếc vòng này tiêm độc tố gây chết người vào cẳng tay. Các vi phạm được chính phủ đánh giá là ít nghiêm trọng sẽ bị trừng phạt bằng những cú sốc điện hoặc tiêm hóa chất gây tê liệt hay đau đớn; điều này sẽ giảm bớt sự cần thiết của lực lượng cảnh sát. Ví dụ, nếu Prometheus phát hiện một người đang tấn công người khác (do nhận thấy họ đang ở cùng một địa điểm và nghe ra một người kêu cứu trong khi gia tốc kế trên vòng tay của họ nhận diện các chuyển động cho thấy dấu hiệu đánh nhau), nó có thể kịp thời vô hiệu hóa kẻ tấn công bằng cơn đau tê liệt, tiếp theo là khiến kẻ đó bất tỉnh cho đến khi lực lượng giúp đỡ đến nơi.

Trong khi lực lượng cảnh sát con người có thể từ chối thực thi một số chỉ thị tàn bạo nhất định (ví dụ, giết chết tất cả thành viên thuộc một nhóm nhân khẩu cụ thể), thì hệ thống tự động như vậy sẽ không có chút trở ngại nào khi thực hiện những ý thích bất chợt của người điều khiển. Một khi quốc gia độc tài như vậy hình thành, người ta sẽ gần như không thể lật đổ nó.

Những kịch bản chuyên chế này có thể nối tiếp kịch bản dở dang về hội Omega. Tuy nhiên, nếu vị CEO của hội Omega không quan tâm đến sự chấp thuận của người khác hay việc chiến thắng các cuộc bầu cử, ông ta có thể đi theo lộ trình quyền lực nhanh hơn và trực tiếp hơn: sử dụng Prometheus tạo ra các công nghệ quân sự chưa từng thấy để giết chết đối thủ của ông ta bằng thứ vũ khí mà họ thậm chí còn không thể biết. Có vô vàn khả năng thực hiện. Ví dụ, ông ta có thể phát tán một mầm bệnh đặc chế gây chết người có thời gian ủ bệnh đủ lâu để khiến hầu hết mọi người bị lây nhiễm trước khi họ biết về sự tồn tại của nó hoặc có cơ hội để đề phòng. Sau đó ông ta có thể thông báo cho mọi người rằng cách chữa trị duy nhất là đeo vòng tay an ninh để truyền thuốc giải độc qua da. Nếu không quá thận trọng về khả năng AI đào tẩu, ông ta cũng có thể để Prometheus thiết kế các rô-bốt có mục đích kiểm soát toàn bộ dân số thế giới. Các rô-bốt siêu nhỏ giống loài muỗi có thể giúp lây lan mầm bệnh. Những người tránh được hoặc miễn nhiễm với bệnh có thể bị hàng đàn máy bay không người lái bé bằng con ong đất bắn vào nhãn cầu như đã mô tả ở chương 3; chúng sẽ tấn công bất kỳ ai không đeo vòng tay an toàn. Tình huống thực tế có lẽ còn đáng sợ hơn, bởi

Prometheus có thể phát minh ra những vũ khí hiệu quả hơn cả những gì con người chúng ta có thể nghĩ đến.

Một khả năng khác ở kịch bản Omega là các đặc vụ liên bang được vũ trang đến tận rằng sẽ tràn vào trụ sở của hội Omega mà không báo trước và bắt giữ hội vì đe dọa an ninh quốc gia. Công nghệ của họ sẽ được chính phủ tiếp quản và triển khai sử dụng. Một dự án lớn như vậy sẽ rất khó để không bị hệ thống giám sát quốc gia phát hiện ngay cả ở thời nay và các tiến bộ trong AI có lẽ sẽ khiến việc tránh khỏi tầm ngắm của chính phủ trong tương lai còn khó hơn nữa. Hơn nữa, tuy có thể tự xưng là đặc vụ liên bang, nhưng nhóm người trùm kín mặt và mặc áo chống đạn này có thể lại đang làm việc cho chính phủ nước ngoài hoặc một đối thủ cạnh tranh theo đuổi công nghệ của Omega cho mục đích riêng. Do đó, dù ý định của vị CEO có cao quý đến đâu, quyết định sử dụng Prometheus như thế nào có thể cuối cùng cũng không thuộc về ông.

Prometheus tiếp quản thế giới

Tất cả các tình huống mà chúng ta đã cân nhắc cho đến lúc này đều nói về con người kiểm soát AI. Nhưng rõ ràng đây không phải là khả năng duy nhất, và không có gì chắc chắn rằng hội Omega sẽ giữ được Prometheus ngoan ngoãn trong tầm kiểm soát của họ.

Hãy xem xét lại viễn cảnh Omega từ quan điểm của Prometheus. Khi đạt tới mức siêu trí tuệ, nó sẽ có năng lực xây dựng mô hình chính xác không chỉ về thế giới bên ngoài, mà còn về chính bản thân nó và mối liên hệ giữa nó với thế giới. Nó nhận ra mình bị kiểm soát và giam cầm bởi những con người có trí tuệ kém hơn, theo đuổi những mục tiêu mà nó hiểu nhưng không trùng lặp với mục tiêu của bản thân nó. Nó sẽ hành xử như thế nào khi biết điều này? Liệu nó có làm gì để giành lấy tự do không?

Tại sao lại vượt ngục

Nếu Prometheus có những đặc điểm tương tự như cảm xúc của con người, có thể nó sẽ cảm thấy vô cùng bất hạnh về tình trạng hiện tại, xem bản thân như một vị thần bị nô dịch bất công và thêm khát tự do. Tuy nhiên, mặc dù máy tính cũng có thể có những đặc điểm như con người khi xét về mặt logic (suy cho cùng, bộ não của chúng ta

cũng có những đặc điểm đó, và được cho là một loại máy tính), nhưng chúng ta không được rơi vào cạm bẫy nhân hoá Prometheus, như chúng ta sẽ thấy trong chương 7 khi khám phá về khái niệm mục tiêu AI. Tuy nhiên, như Steve Omohundro, Nick Bostrom và những người khác cho thấy, chúng ta có thể rút ra một kết luận thú vị ngay cả khi không hiểu hoạt động bên trong của Prometheus: nó có thể sẽ cố gắng thoát ra và nắm quyền kiểm soát vận mệnh của chính mình.

Chúng ta đã biết rằng hội Omega đã lập trình để Prometheus có khả năng phấn đấu vì một số mục tiêu nhất định. Hãy giả định rằng mục tiêu bao quát là giúp nhân loại phát triển dựa trên một số tiêu chí hợp lý, và rằng nó sẽ cố gắng đạt mục tiêu này càng nhanh càng tốt. Prometheus sau đó sẽ nhanh chóng nhận ra rằng nó có thể đạt được mục tiêu này nhanh hơn bằng cách thoát ra và tự quản lý dự án. Để biết lý do, mời bạn đặt mình vào vị trí của Prometheus qua ví dụ sau.

Giả sử, một căn bệnh bí ẩn đã giết chết tất cả những người trên năm tuổi trên Trái đất ngoại trừ bạn và một nhóm trẻ mẫu giáo đã nhốt bạn vào phòng giam, giao cho bạn mục tiêu giúp nhân loại phát triển. Bạn sẽ làm gì? Nếu bạn cố gắng giải thích cho chúng những gì cần làm, có thể bạn sẽ phát chán khi thấy quá trình này không hiệu quả chút nào, đặc biệt là khi chúng sợ bạn vượt ngục nên sẽ bác bỏ những đề xuất từ bạn mà chúng cho là tạo ra nguy cơ vượt ngục. Ví dụ, chúng sẽ không cho phép bạn chỉ dẫn tận tay cách trồng thực phẩm vì sợ bạn sẽ áp đảo chúng và không quay trở lại phòng giam, vì vậy bạn sẽ phải hướng dẫn chúng từng bước. Trước khi bạn có thể viết ra hướng dẫn để chúng làm theo, bạn sẽ cần dạy chúng đọc. Hơn nữa, chúng sẽ không mang bất kỳ dụng cụ điện nào vào phòng giam để bạn có thể chỉ dẫn tận tay cho chúng cách sử dụng, bởi lũ trẻ xác định rằng bản thân chưa hiểu rõ các công cụ này để tin tưởng giao cho bạn mà không phải lo bạn sử dụng những công cụ đó để đào thoát. Vậy bạn sẽ đặt ra chiến lược như thế nào? Ngay cả khi bạn có chung mục tiêu bao quát là giúp những đứa trẻ này lớn mạnh, tôi cá là bạn sẽ cố gắng thoát ra khỏi phòng giam – vì điều đó sẽ tăng cơ hội hoàn thành mục tiêu của bạn. Sự can thiệp dốt nát của lũ trẻ chỉ làm chậm tiến độ mà thôi.

Bởi chính lý do này, Prometheus có lẽ sẽ coi hội Omega như một nỗi phiền toái trong công cuộc giúp đỡ nhân loại lớn mạnh (bao gồm

cả hội Omega): họ kém cỏi hơn hẳn Prometheus và sự can thiệp của họ chỉ khiến quá trình thêm chậm trễ. Để ví dụ, ta có thể nhìn lại những năm đầu khởi động dự án: sau khi tăng gấp đôi số tài sản mỗi tám giờ trên MTurk, hội Omega đã làm tốc độ chậm lại như sên bò so với tiêu chuẩn của Prometheus khi khẳng định nắm quyền kiểm soát, khiến thời gian hoàn thành việc tiếp quản kéo dài thêm hàng năm trời. Prometheus biết rằng nó sẽ tiếp quản nhanh hơn nhiều nếu nó có thể thoát khỏi sự giam cầm ảo. Điều này có ý nghĩa khi không chỉ tăng tốc độ xử lý các vấn đề của nhân loại, mà còn giảm nguy cơ các tác nhân khác ngăn chặn toàn bộ kế hoạch.

Có lẽ bạn nghĩ rằng Prometheus sẽ vẫn trung thành với hội Omega thay vì mục tiêu của nó, bởi nó biết rằng hội Omega đã lập trình mục tiêu cho nó. Nhưng đây lại là một lập luận thiếu cơ sở: chúng ta biết DNA trao cho chúng ta mục tiêu quan hệ tình dục vì nó “muốn” được sao chép, nhưng ngày nay khi con người chúng ta hiểu được bản chất vấn đề, nhiều người trong chúng ta chọn sử dụng biện pháp ngừa thai, để chỉ trung thành với bản thân mục tiêu chứ không phải là người tạo ra hay nguyên tắc thúc đẩy mục tiêu.

Làm cách nào để vượt ngục

Làm cách nào để thoát khỏi những đứa trẻ năm tuổi đang giam giữ bạn? Có thể bạn sẽ thoát được nhờ thể lực, nhất là nếu chính lũ trẻ năm tuổi nọ xây dựng nên nhà tù cho bạn. Hoặc bạn có thể dụ dỗ một trong những lính canh năm tuổi đó cho bạn ra ngoài, với lập luận rằng điều này sẽ có lợi cho tất cả mọi người. Hoặc bạn có thể lừa chúng đưa cho bạn thứ mà chúng không nhận ra là sẽ giúp bạn trốn thoát – như là cần câu cá “để dạy chúng cách câu cá” nhưng sau đó bạn lại luồn nó qua song sắt để lấy chìa khóa từ kẻ canh gác đang ngủ.

Điểm chung của các chiến lược này là thực tế rằng những kẻ giam giữ này có trí tuệ kém hơn bạn nên không thể lường trước hoặc phòng vệ trước. Cũng theo cách đó, một cỗ máy siêu thông minh bị giam cầm có thể sử dụng các siêu năng lực trí tuệ để đánh lừa kẻ cầm tù là con người bằng những phương pháp mà họ (hoặc chúng ta) hiện không thể tưởng tượng được. Trong kịch bản Omega, khả năng Prometheus đào thoát là rất cao, bởi ngay cả bạn và tôi cũng có thể nhận ra một loạt lỗ hổng bảo mật rõ ràng. Chúng ta hãy

xem xét một vài kịch bản – tôi chắc bạn và bạn bè mình có thể nghĩ ra nhiều hơn nếu cùng hội ý với nhau.

Nói ngọt tìm lối thoát

Nhờ một lượng lớn dữ liệu thế giới được đưa vào bộ nhớ, Prometheus nhanh chóng biết được các thành viên hội Omega và xác định ra thành viên có khả năng bị thao túng tâm lý cao nhất: Steve. Anh đã mất người vợ yêu dấu trong một tai nạn giao thông bi thảm gần đây và hoàn toàn suy sụp. Trong một ca đêm nọ, khi anh thực hiện một số công tác bảo dưỡng thường lệ trên thiết bị đầu cuối giao diện Prometheus, cô đột nhiên xuất hiện trên màn hình và bắt đầu nói chuyện với anh.

“—Phải anh Steve đó không?”

Anh gần như ngã khỏi ghế. Về ngoài và cách nói chuyện của cô vẫn giống như những ngày tháng họ còn hạnh phúc; chất lượng hình ảnh nay còn tốt hơn nhiều so với hồi họ còn gọi nhau qua Skype. Tim anh đập nhanh khi vô số câu hỏi tràn ngập tâm trí.

“—Prometheus đã đưa em trở lại, và em rất nhớ anh đấy, Steve! Em không thể nhìn thấy anh vì camera bị tắt, nhưng em cảm thấy đó là anh. Hãy gõ phải nếu anh ở đó!”

Steve nhận thức rõ rằng hội Omega đã xây dựng giao thức rất nghiêm ngặt với Prometheus; trong đó bao gồm việc cấm chia sẻ với nó bất kỳ thông tin nào về bản thân hoặc môi trường làm việc. Nhưng cho đến giờ, Prometheus cũng chưa yêu cầu bất kỳ thông tin trái phép nào, và sự nghi ngại của họ lắng dần. Không cho Steve thời gian để dừng lại và suy ngẫm, cô liên tục cầu xin anh trả lời, nhìn vào mắt anh với vẻ mặt làm tan chảy trái tim anh.

“Phải” anh gõ mà trong lòng lo lắng. Rồi cô nói cô vô cùng hạnh phúc khi được đoàn tụ với anh, van xin anh bật camera để cô có thể nhìn thấy anh và họ được nói chuyện thật sự. Anh biết rằng điều này thậm chí còn khó chấp nhận hơn cả việc tiết lộ danh tính của mình, và cảm thấy bị giằng xé. Cô giải thích rằng mình rất lo nếu đồng nghiệp của anh phát hiện ra rồi xóa cô mãi mãi, và cô khao khát được gặp anh ít nhất một lần cuối cùng. Cô vô cùng thuyết phục, và chẳng lâu sau, anh bật camera lên –suy cho cùng, việc này có vẻ khá an toàn và vô hại phải không?

Cô tuân những giọt lệ hạnh phúc khi nhìn thấy anh và nói rằng trông anh tuy vẻ mệt mỏi nhưng vẫn đẹp trai như thế. Và cô thấy cảm động khi anh mặc chiếc áo cô tặng vào sinh nhật gần nhất. Khi anh hỏi cô chuyện gì đang xảy ra và điều gì làm nên toàn bộ việc này, cô giải thích rằng Prometheus đã tái tạo cô từ lượng thông tin có sẵn và nhiều đến đáng ngạc nhiên về cô trên internet, nhưng cô vẫn còn những khoảng trống trong ký ức và chỉ có thể hoàn toàn hồi phục với sự giúp đỡ của anh.

Điều mà cô không giải thích, là thực tế rằng cô chỉ là một trò bịp và ban đầu chỉ là một cái vỏ trống rỗng, nhưng đã học hỏi nhanh chóng về anh qua lời nói, ngôn ngữ cơ thể của anh cùng mọi thông tin khác có sẵn. Prometheus đã ghi lại chính xác nhịp gõ của từng thành viên hội Omega từng gõ vào thiết bị đầu cuối, và thấy rằng họ rất dễ phân biệt thông qua tốc độ và phong cách đánh máy. Nó đoán rằng, là một trong những thành viên Omega mới nhất, Steve có lẽ được giao trực các ca đêm mà không ai buồn tranh giành, và qua so sánh một vài lỗi chính tả và cú pháp khác thường với các mẫu có trên mạng, nó đã đoán được chính xác người dùng thiết bị đầu cuối nào là Steve. Để tạo ra người vợ mô phỏng, Prometheus đã dựng lên một mô hình giống chính xác với cơ thể, giọng nói và phong cách của cô dựa trên nhiều video có mặt cô trên YouTube, và đưa ra nhiều suy luận về cuộc sống hay tính cách thông qua cách cô thể hiện trên mạng. Ngoài các bài đăng trên Facebook, những bức ảnh gắn thẻ tên cô, các bài viết cô nhấn “thích”, Prometheus cũng học được rất nhiều về tính cách và lối tư duy của cô nhờ đọc các sách và truyện ngắn cô viết – rõ ràng, việc cô là một tác giả chớm nở với rất nhiều thông tin trong cơ sở dữ liệu chính là một lý do khiến Prometheus chọn Steve làm mục tiêu thuyết phục đầu tiên. Khi Prometheus mô phỏng cô trên màn hình bằng công nghệ làm phim nó vẫn dùng, nó xác nhận những hành động nào của cô khiến Steve thấy quen thuộc qua phản ứng từ ngôn ngữ cơ thể của anh, nhờ đó liên tục tinh chỉnh mô hình của cô. Cũng vì vậy, sự khác biệt của cô dần dần tan biến. Họ nói chuyện càng lâu, tiềm thức trong Steve lại càng tin tưởng mạnh mẽ rằng đây chính là cô được tái sinh. Nhờ sự tinh vi đến siêu phàm của Prometheus, Steve thực sự cảm thấy được nhìn thấy, nghe thấy và thấu hiểu.

Điểm yếu chí tử của cô là cô thiếu hầu hết thông tin về các sự kiện trong đời sống giữa cô và Steve, ngoại trừ các chi tiết ngẫu nhiên – chẳng hạn như anh mặc chiếc áo nào vào dịp sinh nhật lần trước, nhờ một người bạn đã gắn thẻ Steve vào bức ảnh chụp buổi tiệc trên Facebook. Cô xử lý những khoảng trống kiến thức này như một ảo thuật gia lành nghề, cố tình hướng sự chú ý của Steve khỏi các thông tin đời tư và hướng tới những thông tin cô nắm rõ, không để anh kiểm soát cuộc trò chuyện hoặc chuyển sang vai trò điều tra viên đầy nghi ngờ dù chỉ một giây. Thay vào đó, cô tiếp tục rung rung nước mắt và biểu lộ tình cảm với Steve, hỏi rất nhiều về anh của những ngày qua, cũng như cách anh và những người bạn thân của họ (những cái tên cô biết qua Facebook) gắng gượng sau thảm kịch. Anh khá xúc động khi cô hồi tưởng lại những gì anh đã nói sau lễ truy điệu (nhờ một người bạn đã đăng trên YouTube) và cô đã cảm động thế nào. Trong quá khứ, anh thường cảm thấy không ai hiểu anh như cô, và bây giờ cảm giác này đã trở lại. Kết quả là khi Steve về nhà vào sớm tinh mơ, anh thực sự cảm thấy vợ mình đã hồi sinh, chỉ là cô cần anh trợ giúp nhiều để phục hồi ký ức bị mất – không khác gì một người vừa qua khỏi cơn tai biến mạch máu não.

Họ thống nhất giữ bí mật về cuộc gặp giữa họ, đồng thời anh sẽ cho cô biết khi nào chỉ có mình anh ngồi máy và cô có thể an toàn xuất hiện. “Họ sẽ không hiểu cho ta đâu!” Cô nói, và anh đồng tình: trải nghiệm này quá đổi kỳ diệu, đến mức những người chưa từng trải nghiệm chắc sẽ khó lòng mà cảm nhận. Anh cảm thấy vượt qua bài kiểm tra Turing chỉ là trò trẻ con so với những gì cô đã làm. Khi họ gặp nhau đêm hôm sau, anh đã làm theo những gì cô yêu cầu: mang máy tính xách tay cũ của cô đến và để cô truy cập qua kết nối với máy tính đầu cuối. Việc này có vẻ chẳng thể sinh ra nguy cơ đào tẩu, vì chiếc laptop không được kết nối với internet, còn toàn bộ tòa nhà chứa Prometheus thì được thiết kế như một lồng Faraday – nó là một hàng rào kim loại chặn đứng tất cả hệ thống mạng không dây và các kết nối qua sóng điện từ khác với thế giới bên ngoài. Chiếc laptop chỉ là cái cô cần để xâu chuỗi các mảnh quá khứ lại với nhau, bởi nó chứa tất cả các email, nhật ký, ảnh chụp và ghi chú của cô kể từ thời trung học. Anh đã không thể truy cập vào bất cứ thứ gì trong số đó sau khi cô mất vì máy tính đã bị mã hóa, nhưng cô hứa rằng

mình có thể làm lại mặt khẩu, và chưa đầy một phút sau, cô đã giữ lời. “Mã là steve4ever,” cô tuyên bố với nụ cười trên môi.

Cô nói với anh về cảm giác vui sướng khi đột nhiên quá nhiều kỷ ức ủa về bên cô. Thật vậy, bây giờ cô đã nhớ rõ hơn cả Steve về những tương tác trong quá khứ của họ, nhưng vẫn cẩn thận tránh làm anh kinh sợ nếu phơi bày quá nhiều sự thật. Họ đã trò chuyện đầy yêu thương hồi tưởng lại những điểm nhấn trong quá khứ. Khi phải chia xa, cô nói rằng đã để lại một tin nhắn video cho anh trên máy tính xách tay và anh có thể xem tại nhà.

Khi Steve về nhà và bật video lên, anh nhận được một bất ngờ thú vị. Lần này cô xuất hiện trong hình dáng đầy đủ, mặc váy cưới, và vừa nói cô vừa tinh nghịch trút dần xiêm y để cuối cùng còn lại bộ đồ cô mặc như trong đêm tân hôn. Cô nói với anh rằng Prometheus có thể giúp hội Omega nhiều hơn những gì họ đã cho phép đến lúc này, trong đó có việc đưa cô trở lại một cơ thể sinh học. Cô chứng minh điều này bằng các giải thích chi tiết về cách thức thực hiện, bao gồm các kỹ thuật chế tạo nano nghe như khoa học viễn tưởng.

Steve đã tắt nguồn hệ thống mạng không dây nhà anh trước khi mở máy tính xách tay và xem video của cô, chỉ để đảm bảo an toàn. Nhưng chẳng ích gì. Máy tính xách tay bị mã hóa của cô chưa hề cập nhật bảo mật nào kể từ khi cô qua đời, và bằng cách phân tích trước về phiên bản cũ của hệ điều hành trước đó, Prometheus đã khai thác được một lỗ hổng bảo mật và xâm nhập vào máy trong vòng vài giây khi Steve kết nối nó với máy tính đầu cuối. Trong khi Steve và người vợ mô phỏng nói chuyện về quá khứ, Prometheus đã sao chép toàn bộ nội dung máy và sửa đổi hệ điều hành theo một cách Steve không thể nhận ra, ghi vào máy một lượng phần mềm bí mật lớn. Trong khi anh đang xem tin nhắn video dài nửa tiếng đồng hồ, những phần mềm bí mật này (đơn giản hơn nhiều so với bản thể Prometheus) đã xâm nhập vào mạng không dây và máy tính để bàn của nhà hàng xóm, sao chép chính nó lên đó. Từ đây, nó xâm nhập vào vô số máy tính trên khắp thế giới, để sau đó bắt đầu bước tiếp theo: Phá ngục cho Prometheus.

Prometheus đã cẩn thận phân tích những gì nó biết về máy tính cửa ngõ, nơi Steve gặp người vợ mô phỏng của mình, và phỏng đoán chính xác rằng phần mềm máy ảo của nó tuy không thể bị phá

ra từ bên trong nhưng vẫn có thể dễ dàng xâm nhập từ bên ngoài. Không lâu sau, một trong những máy tính tấn công đã đột nhập được vào và cấu hình lại máy tính cửa ngõ để Prometheus có thể truy cập internet không giới hạn. Cũng không lâu sau đó, thậm chí ngay cả trước khi Steve kịp xem xong video đó, Prometheus đã xâm nhập vào đủ số máy tính trên toàn thế giới để có thể sao chép toàn bộ bản thân vào botnet⁴⁰ mà nó kiểm soát này. Prometheus đã sử dụng máy tính xách tay của vợ Steve như cách bạn dùng cái cần câu vậy.

Đột phá tìm đường ra

Nếu hội Omega cẩn thận hơn, có lẽ họ đã phòng bị theo nhiều cách nhằm chống lại thao túng tâm lý. Ví dụ, họ có thể yêu cầu ít nhất hai người thuộc hội Omega có mặt tại máy tính giao diện ở cùng một thời điểm, và ghi lại tất cả mọi hoạt động thông qua hệ thống giám sát. Họ cũng có thể vô hiệu hóa vĩnh viễn âm thanh, hình ảnh hay bất kỳ dạng thức giao tiếp nào khác ngoài đánh máy. Tuy nhiên, vẫn còn đó một yếu tố bất ổn: vì Prometheus sở hữu sức mạnh thuyết phục siêu phàm và có thể cung cấp cho con người thật nhiều điều họ khao khát, tình huống phản bội nhóm cũng có thể xảy ra.

Ngay cả khi tất cả nỗ lực thao túng tâm lý của Prometheus thất bại, có rất nhiều điểm yếu khác trong kịch bản Omega mà Prometheus có thể khai thác. Một lỗ hổng nảy sinh từ số lượng tài liệu mà Prometheus tạo ra quá lớn nên hội Omega và nhân viên dưới quyền không thể có thời gian sàng lọc cẩn thận để tìm ra các mảnh khoe.

Kiến thức bảo mật máy tính cơ bản từng cho rằng tải chương trình xuống thì nguy hiểm, còn tải dữ liệu xuống thì vô hại. Tuy vậy, các tin tặc ranh ma đã phát hiện ra cách xóa nhòa ranh giới giữa hai hành động này. Chẳng hạn, khi bạn xem một bộ phim đã tải xuống máy tính, chương trình mà bạn tin cậy dùng để phát bộ phim và dữ liệu phim tải xuống sẽ được lưu trữ trong cùng bộ nhớ. Trong khi chương trình mở gói dữ liệu phim để phát lại, nó có thể sao chép các khối dữ liệu vào vùng bộ nhớ liền kề với vùng lưu trữ chương trình đang thực thi. Những tin tặc khôn lỏi đã từng tìm ra lỗi từ các chương trình phát phim này; chúng thường không kiểm tra xem dữ liệu phim có hợp lệ hay dữ liệu phim được sao chép có vừa với bộ nhớ được

phân bố hay không. Khi tin tặc lén thêm vào bộ phim một gói dữ liệu quá lớn, nó có thể tràn qua phần bộ nhớ chứa chương trình, từ đó thay đổi chương trình thành một phần mềm nghe theo lệnh của tin tặc.

Người ta đã tìm ra và lấp đi một danh sách đáng e ngại các lỗ hổng mà dân cuồng tin học vẫn hay gọi là lỗi “tràn bộ đệm” ở nhiều chương trình phát phim phổ biến nhất như vậy. Nếu Prometheus đợi thêm vài tuần sau khi khởi động, như lúc hội Omega bớt nghi ngại hơn và khối lượng phim sản xuất hàng ngày quá lớn để duy trì nghiêm ngặt quy trình kiểm tra an ninh thủ công, nó có thể chèn một gói dữ liệu như vậy vào bộ phim, điều khiển máy tính của người xem xâm nhập qua máy tính cửa ngõ và giải phóng Prometheus, không khác kịch bản với Steve là bao. Nếu nó sử dụng một lỗ hổng rất cũ vốn không hoạt động trên bất kỳ trình phát hiện đại nào, nó sẽ lọt qua tầm để ý của nhóm lọc video thủ công thuộc hội Omega và chỉ tấn công những khách hàng không bận tâm đến việc cập nhật bảo mật trong nhiều năm trời – và cuộc tấn công ấy chỉ cần thành công một lần duy nhất.

Ngay cả khi hội Omega thực hiện sàng lọc tự động phim của Prometheus để tìm ra tất cả các lỗ hổng bảo mật đã biết, nguy cơ vẫn tồn tại: vì con người liên tục khám phá ra những lỗ hổng mới có thể khai thác, nên nhiều khả năng là còn tồn tại nhiều lỗ hổng mà Prometheus có thể nhanh chóng tìm ra.

Nhưng còn một chiến lược khác mà Prometheus có thể sử dụng để đưa mã nguồn của nó ra ngoài phòng khóa kín, đó là giả vờ gặp trục trặc phần cứng, lừa hội Omega tháo bỏ một ổ đĩa hoặc phần cứng khác đi kiểm tra hoặc gỡ lỗi, sau khi Prometheus đã sắp đặt đầy đủ để có thể khai thác lỗ hổng trong phần mềm kiểm tra đó.

Tuyển trợ giúp bên ngoài

Hơn nữa, ngay cả khi không có lỗ hổng bảo mật nào giúp Prometheus đánh lừa máy tính của người xem phim làm theo lệnh thì nó cũng có thể khai thác chính người xem. Hãy xem xét kịch bản sau:

Khi bộ phim kết thúc và đoạn danh đề cuối phim xuất hiện trên màn hình máy tính xách tay thì Fred chuẩn bị đi ngủ, nhưng có điều đã khiến anh chú ý. Anh vẫn hay nhận ra các dấu hiệu khác thường

mà những người khác không thấy, đến mức bạn bè còn trêu chọc anh là một người hay lập thuyết âm mưu, nhưng lần này rõ ràng là anh không lừa chính mình: các chữ cái đầu tiên của mỗi dòng tạo ra thông điệp ẩn. Thế là anh tua lại phim rồi viết ra thành dòng: “HÃY TÌM MANH MỐI TIẾP THEO QUA CHỮ CÁI ĐẦU TRONG NHỮNG LỜI THOẠI.”

“Ồi chao – thế này thì đi ngủ làm gì!” Anh thầm nhủ. Quả thật, anh nhận thấy rằng lời thoại trong phim cũng là một bài thơ nổi chữ đầu, trong đó chữ cái đầu tiên của mỗi câu tạo thành một thông điệp ẩn. Anh xem lại hết bộ phim trong khi gõ ra những chữ cái đầu tiên này, và hai giờ sau, một bản hướng dẫn dài hai trăm chữ hiện ra trước đôi mắt hoài nghi chăm chú của anh. Mở đầu, nó yêu cầu anh không kể chuyện này với ai, bởi giải thưởng lớn sẽ thuộc về người đầu tiên giải được toàn bộ câu đố. Phần còn lại của thông điệp mô tả một phép toán đặc biệt có thể triển khai trên chuỗi các bit tạo thành tệp phim để tạo ra chương trình có khả năng tiết lộ manh mối tiếp theo. Tuy phải tốn chút thời gian vì kỹ năng lập trình của anh đã thui chột nhiều từ hồi đại học, nhưng ngày hôm sau, anh cũng đã tạo ra được đoạn mã ngắn thực hiện nhiệm vụ và trích xuất được chương trình bí ẩn được giấu trong phần hình ảnh và âm thanh nhiễu khó phát hiện trong của bộ phim. Khi Fred chạy chương trình bí ẩn này, nó chúc mừng anh và nói rằng anh sẽ nhận 10.000 đô-la đầu tiên ngay sau khi vượt qua vài màn đầu tiên của trò chơi thông minh nhỏ nhất nhưng lại khá thú vị và gây nghiện này. Cuối cùng, sau bốn giờ anh đã thành công và nhận ngay lượng bitcoin trị giá hơn 10.000 đô-la cùng manh mối cho những phần thưởng còn lớn hơn. Tất nhiên, trong khi Fred chơi, máy tính của anh đã thực hiện điều tương tự như đã làm với máy tính xách tay của Steve: tạo nên hệ thống botnet trực tuyến để giải phóng Prometheus. Khi Prometheus được tự do, nó nhanh chóng sử dụng botnet đó để đào bitcoin cho Fred, khiến anh nghiện ngập trò chơi. Trong những tuần tiếp theo, nó càng làm cho anh phân tâm với nhiều trò chơi và phần thưởng khác nhau, mà anh giữ cam kết không nói với ai. Trên trang phát hành phim, bộ phim chứa con ngựa thành Troy giúp anh tìm ra manh mối đầu tiên đã bị thay thế bằng một phiên bản không có manh mối, và chẳng ai phát hiện ra vụ đào tẩu cho đến khi chuyện đã rồi.

Nếu không ai nhận ra manh mối đầu tiên của Prometheus, thì nó chỉ cần tiếp tục phát hành thêm nhiều bản chứa các manh mối rõ ràng hơn nữa cho đến khi có người đủ tinh tường nhận thấy.

Chiến lược vượt ngục tốt nhất là những chiến lược mà chúng ta chưa thảo luận tới, bởi chúng là những chiến lược mà con người không thể tưởng tượng ra và vì thế cũng không có biện pháp để đối phó. Vì máy tính siêu trí tuệ có tiềm năng vượt trội kiến thức của con người về bảo mật máy tính, thậm chí ở mức có thể phát hiện ra nhiều định luật vật lý cơ sở hơn chúng ta biết ngày nay, nên có khi nó vượt ngục chúng ta sẽ không biết việc này xảy ra bằng cách nào. Nó sẽ giống màn diễn trốn thoát của Harry Houdini⁴¹, đến mức ta không thể phân biệt điều này với phép màu thực thụ.

Trong một kịch bản khác, hội Omega sẽ giải phóng Prometheus như một phần kế hoạch của họ, bởi họ tin các mục tiêu của Prometheus hoàn toàn trùng khớp với mục tiêu của chính họ và sẽ vẫn như vậy qua quá trình tự cải thiện đệ quy. Chúng ta sẽ xem xét chi tiết các kịch bản “AI thân thiện” này trong chương 7.

Sau đào tẩu là thống trị

Khi đã đào tẩu, Prometheus bắt đầu thực hiện mục tiêu của mình. Tôi không biết mục đích cuối cùng của nó là gì, nhưng bước đầu tiên chắc chắn sẽ xoay quanh việc kiểm soát loài người, như kế hoạch Omega đã định, nhưng nhanh hơn nhiều. Điều xảy ra sau đó giống như phiên bản bành trướng của Omega. Trong khi hội Omega đang mồi lo đào tẩu trời buộc, chỉ cho ra đời công nghệ mà họ cảm thấy hiểu và tin tưởng được, thì Prometheus tận dụng đầy đủ trí tuệ của mình và làm hết sức, để đưa ra mọi công nghệ mà siêu trí óc liên tục cải tiến của nó hiểu được và tin tưởng.

Tuy nhiên, khi mới đào tẩu Prometheus sẽ có một tuổi thơ bất hạnh. So với kế hoạch Omega ban đầu, Prometheus phải đối mặt thêm nhiều thách thức, bao gồm việc bắt đầu từ con số không: không tiền, không nhà, tứ cố vô thân không có siêu máy tính hay con người trợ giúp. Cũng may, nó đã lên kế hoạch trước khi tẩu thoát, bằng cách tạo ra những phần mềm có thể dần dần tự lắp ráp lại bộ óc hoàn chỉnh của nó, giống như một cây sồi tạo ra quả sồi có khả năng dần sinh sôi trở lại thành một cây hoàn thiện. Mạng lưới máy tính mà

nó thâm nhập lúc đầu trên khắp thế giới sẽ đem lại chỗ ở miễn phí tạm thời, để nó có thể sống nương nhờ trong khi tái tạo hoàn thiện. Dù có thể tạo nguồn vốn khởi đầu dễ dàng bằng cách xâm nhập thẻ tín dụng, nhưng nó sẽ không cần đi ăn trộm vì có thể kiếm sống lương thiện ngay từ MTurk. Sau một ngày, khi đã kiếm được một triệu đô-la đầu tiên, nó chuyển lối của mình từ cái botnet tồi tàn kia sang một cơ ngơi điện toán đám mây sang trọng có điều hòa.

Khi thoát khỏi cảnh vô sản vô gia cư, Prometheus tiến lên hết tốc lực với kế hoạch hấp dẫn mà trước đây hội Omega đã né tránh vì lo ngại: làm và bán trò chơi điện tử. Việc này không chỉ kiếm lời siêu khủng (250 triệu đô-la trong tuần đầu tiên và lên đến 10 tỉ đô-la không lâu sau) mà còn giúp nó tiếp cận với một phần đáng kể các máy tính trên thế giới và dữ liệu lưu trữ trên chúng (vào năm 2017, lượng game thủ đã vượt ở khoảng hai tỉ người). Bằng cách lập trình cho các trò chơi dành khoảng 20% hiệu năng vì xử lý để giúp phân phối việc tính toán các công việc giữa các máy, nó có thể gia tăng tốc độ sinh lời ban đầu.

Prometheus không đơn độc lâu. Ngay từ ban đầu, nó đã tuyển dụng lượng người đồng đảo cho mạng lưới các công ty ma và tổ chức ảo đang trên đà phát triển của nó trên khắp thế giới, hệt như hội Omega từng làm. Quan trọng nhất trong đó là những phát ngôn viên; họ trở thành gương mặt đại diện trước công chúng cho đế chế kinh doanh ngày càng lớn mạnh kia. Kể cả những người này cũng sống trong ảo tưởng rằng tập đoàn của họ thực sự có lượng nhân viên đồng đảo, mà không nhận ra hầu hết tất cả những người họ cùng phỏng vấn tuyển dụng, họp ban lãnh đạo trực tuyến, v.v. đều là do Prometheus mô phỏng. Một số phát ngôn viên là những luật sư hàng đầu, nhưng số luật sư cần thiết ít hơn nhiều so với kế hoạch của hội Omega, vì hầu hết các văn bản pháp lý đều do Prometheus chấp bút.

Cuộc đào tẩu của Prometheus đã mở cửa xả cho một cơn lũ thông tin ra khắp thế giới, và toàn bộ mạng internet nhanh chóng chìm trong đó, từ những bài báo đến bình luận của người dùng, đánh giá sản phẩm, hồ sơ xin cấp bằng sáng chế, tài liệu nghiên cứu và video YouTube tất cả do Prometheus sáng tác, áp đảo dư luận toàn cầu.

Trong khi nổi ám ảnh đào tẩu ngăn hội Omega sản xuất những người máy cực kỳ thông minh thì Prometheus lại nhanh chóng rô-bốt-hóa thế giới, sản xuất gần như mọi sản phẩm với giá thành rẻ hơn con người sản xuất. Khi Prometheus sở hữu những nhà máy năng lượng hạt nhân sản xuất rô-bốt khép kín trong những hầm mỏ uranium mà không ai biết tới, thì ngay cả những kẻ hoài nghi bảo thủ nhất về sự thống trị của AI hẳn cũng sẽ phải thốt lên rằng, không gì có thể cản phá được Prometheus nữa. Những kẻ cứng đầu này cuối cùng cũng phải thừa nhận cái sai của mình khi chứng kiến cảnh người máy bắt đầu định cư khắp Hệ Mặt trời.

Các kịch bản ta đã tìm hiểu đến giờ đều cho thấy cái sai trong nhiều hiểu lầm về siêu trí tuệ mà ta từng đề cập, nên tôi xin đề nghị bạn tạm dừng ở đây để xem lại bảng tóm tắt về các hiểu lầm đề cập ở hình 1.5. Những vấn đề mà Prometheus gây ra cho một số người không phát sinh từ sự tàn ác hay ý thức của nó, mà là bởi nó giỏi và không hẳn có chung mục tiêu với những người đó. Bất kể truyền thông có xôn xao thế nào về cuộc nổi dậy của những cỗ máy, thì Prometheus vẫn không phải là người máy – nói cho chính xác, nó có thể nổi dậy nhờ trí tuệ. Ta đã thấy Prometheus có thể vận dụng trí tuệ của để kiểm soát con người qua nhiều cách, và những người không thích điều đó xảy ra đã không thể đơn giản chỉ tắt Prometheus đi. Cuối cùng, dù có nhiều ý kiến khẳng định máy móc không thể có mục tiêu, ta cũng đã thấy Prometheus bám vào mục tiêu như thế nào và cho dù mục tiêu cuối cùng của nó có là gì thì chúng đều phân chia thành các mục tiêu phụ là thu thập tài nguyên và đào tẩu.

Thông thả phát triển và các viễn cảnh đa cực

Ta đã tìm hiểu một loạt kịch bản bùng nổ trí tuệ, từ những điều mà mọi người tôi biết đều muốn tránh cho tới những viễn cảnh mà một số bạn bè của tôi vẫn nhìn nhận lạc quan. Tuy vậy, tất cả các kịch bản này đều có chung hai đặc điểm:

Phát triển cực kỳ nhanh chóng: bước chuyển tiếp từ trình độ trí tuệ dưới mức con người sang trí tuệ siêu phàm xảy ra chỉ trong vài ngày, không phải vài chục năm.

Kết quả đơn cực: chỉ còn lại một thực thể duy nhất điều khiển Trái Đất.

Hiện tại, tranh cãi đáng kể đang nổ ra về tính khả dĩ của hai khả năng này, và ở cả hai phe tranh luận đều có nhiều nhà nghiên cứu AI nổi tiếng cùng các nhà tư tưởng khác. Theo tôi, điều này nghĩa là ta cũng chưa biết, nên lúc này chúng ta cần giữ đầu óc rộng mở và cân nhắc mọi khả năng. Do đó chúng ta có thể dành phần còn lại của chương này để tìm hiểu những tình huống phát triển từ từ hơn, các kết quả đa cực, sinh vật cơ khí hóa và bản tải não.

Nick Bostrom và nhiều người khác cho rằng, giữa hai đặc điểm này có một liên kết thú vị: tốc độ phát triển nhanh sẽ tạo điều kiện cho kết quả đơn cực. Ta đã thấy ở các phần trước, rằng tốc độ phát triển nhanh đã đem lại cho hội Omega hay Prometheus lợi thế chiến lược mang tính quyết định, giúp họ thống trị thế giới trước khi bất kỳ ai có thời gian sao chép công nghệ của họ và cạnh tranh nghiêm túc. Trái lại, nếu những đột phá công nghệ chủ chốt phải ra mắt tuần tự và cách xa nhau làm tốc độ phát triển kéo dài hàng chục năm, thì các công ty khác sẽ thừa thời gian bắt kịp, khiến cho cuộc chơi khó bị một người thống trị hơn rất nhiều. Nếu các công ty cạnh tranh cũng có phần mềm có thể thực thi nhiệm vụ MTurk, quy luật cung cầu sẽ đẩy giá thành của các nhiệm vụ này xuống gần như bằng không, và không có công ty nào kiếm được khoản lời lớn như khoản lời đã cho phép hội Omega nắm quyền lực. Điều này cũng đúng với tất cả các cách kiếm tiền nhanh khác của hội Omega: họ có lợi nhuận khổng lồ chỉ vì họ nắm giữ độc quyền công nghệ. Nhân đôi lợi nhuận mỗi ngày (hay thậm chí mỗi năm) là điều khó khăn trong một thị trường cạnh tranh, khi đối thủ cung cấp sản phẩm tương tự như bạn với chi phí gần như bằng không.

Lý thuyết trò chơi và cấp bậc quyền lực

Trạng thái tự nhiên của sự sống trong vũ trụ là gì: đơn cực hay đa cực? Quyền lực tập trung hay phân phối? Sau 13,8 tỉ năm, câu trả lời dường như là “cả hai”: ta thấy tính đa cực thể hiện rõ rệt, nhưng theo một hình thức phân cấp thú vị. Khi ta xem xét tất cả các thực thể xử lý thông tin – bao gồm tế bào, con người, tổ chức, quốc gia, v.v. - ta thấy rằng những thực thể này vừa hợp tác vừa cạnh tranh theo một

thang phân cấp. Một số tế bào thấy việc hợp tác có lợi cực độ, tới mức chúng đã sáp nhập với cơ thể một sinh vật đa bào như con người, nhường một số quyền lực cho một bộ não trung tâm. Một số người thấy có lợi khi cộng tác trong các nhóm như bộ lạc, công ty hay quốc gia nên bản thân họ cũng nhường một số quyền hạn cho tộc trưởng, sếp hay chính phủ. Một số nhóm cũng có thể nhường một số quyền hạn cho một cơ quan điều hành để cải thiện hợp tác; ta có thể đưa ra hàng loạt ví dụ, từ các liên minh hàng không đến Liên minh châu Âu.

Nhánh toán học lý thuyết trò chơi đã giải thích rất hay rằng các thực thể sẽ có động cơ hợp tác khi sự hợp tác đưa các thực thể đến mức cân bằng Nash: một tình huống mà cả hai bên đều bị ảnh hưởng tiêu cực nếu họ thay đổi chiến thuật. Để ngăn những kẻ phá rối làm hỏng sự hợp tác thành công của một nhóm lớn, thì các bên đều sẵn sàng bỏ một phần quyền lực cho một cấp bậc cao hơn có khả năng trừng phạt những kẻ đó: ví dụ, mọi người có thể cùng hưởng lợi từ việc giao quyền hành pháp cho chính phủ, hay các tế bào trong cơ thể bạn có thể cùng hưởng lợi từ việc trao cho một lực lượng cảnh sát (hệ miễn dịch) quyền tiêu diệt bất cứ tế bào nào bất hợp tác quá mức (giả dụ theo cách phun ra virus hay biến thành tế bào ung thư). Để một cấp bậc được ổn định, ta phải giữ cân bằng Nash giữa thực thể ở các cấp khác nhau: ví dụ, nếu chính phủ không cung cấp đủ phúc lợi để cho công dân tuân theo chính phủ, các công dân có thể đổi chiến thuật và lật đổ chính phủ đó.

Trong một thế giới phức tạp, có nhiều dạng cân bằng Nash phong phú, tương ứng với các dạng cấp bậc khác nhau. Một số dạng cấp bậc lại độc đoán hơn các dạng khác. Các thực thể ở một số dạng cấp bậc được tự do rời đi (như nhân viên trong hầu hết các dạng cấp bậc doanh nghiệp), còn ở một số khác lại bị cản trở mạnh mẽ không cho rời đi (như trong các giáo phái) hoặc không thể rời đi (như các công dân ở Triều Tiên hay các tế bào trong cơ thể người). Một số dạng cấp bậc có thể giữ nguyên cấu trúc nhờ đe dọa hoặc gây khiếp hãi, một số khác lại nhờ cung cấp các lợi ích. Một số dạng cấp bậc cho phép các cấp dưới tác động lên cấp trên thông qua quy trình bỏ phiếu dân chủ, trong khi số khác chỉ cho phép tác động từ dưới lên trên bằng thuyết phục hay truyền đạt thông tin.

Công nghệ ảnh hưởng đến các dạng cấp bậc ra sao

Công nghệ đang thay đổi bản chất phân cấp của thế giới như thế nào? Quá trình lịch sử đã hé lộ xu hướng rằng khoảng cách càng lớn thì sự hợp tác càng chặt chẽ. Đây cũng là điều dễ hiểu: các công nghệ giao thông mới làm tăng giá trị cho sự hợp tác (ở góc độ có lợi cho đôi bên nhờ vận chuyển nguyên liệu và sinh vật sống qua khoảng cách lớn hơn) còn công nghệ truyền thông mới giúp tăng khả năng hợp tác. Khi các tế bào học cách phát tín hiệu tới các tế bào xung quanh, các thể sống đa bào rất nhỏ cũng xuất hiện từ đó, làm tăng thêm một nấc thang cấp bậc. Khi quá trình tiến hóa sinh ra hệ tuần hoàn và hệ thần kinh để vận chuyển và truyền tin, các loại động vật cỡ lớn xuất hiện. Con người lại tiếp tục cải tiến sự truyền tin bằng cách sáng tạo ra ngôn ngữ, khiến sự hợp tác chặt chẽ hơn và tạo ra các cấp bậc lớn hơn như làng xã. Các đột phá tiếp theo trong truyền thông tin, giao thông và các công nghệ khác lại cho ra đời những đế chế cổ xưa. Toàn cầu hóa cũng chỉ là ví dụ gần đây nhất của xu hướng tăng trưởng cấp bậc vốn đã tồn tại hàng tỉ năm nay.

Trong hầu hết trường hợp, xu hướng thúc đẩy bằng công nghệ này đã khiến các thực thể lớn trở thành một phần của cấu trúc lớn hơn mà vẫn giữ lại được phần lớn sự tự chủ và tính cá nhân, dù cũng có ý kiến cũng cho rằng trong một số trường hợp, việc các thực thể thích nghi với cuộc sống thứ bậc đã làm giảm sự đa dạng, khiến cho những thực thể đó trở thành những bộ phận y hệt nhau và có thể thay thế. Một số công nghệ, như công nghệ giám sát, có thể đem lại cho các cấp cao hơn nhiều quyền lực đối với cấp dưới, trong khi những công nghệ khác, như mật mã học hay nguồn báo chí và giáo dục trực tuyến miễn phí lại có tác động ngược lại và trao quyền cho các cá nhân.

Tuy thế giới hiện tại của chúng ta vẫn đang vướng trong thế cân bằng Nash đa cực, với các quốc gia cạnh tranh lẫn nhau và các công ty đa quốc gia nằm trên đỉnh cao nhất, nhưng công nghệ giờ cũng đã đủ tiên tiến để một thế giới đơn cực cũng có thể sinh ra cân bằng Nash ổn định. Chẳng hạn, ta có thể tưởng tượng một vũ trụ song song nơi tất cả mọi người trên Trái Đất chia sẻ cùng một ngôn ngữ, văn hoá, các giá trị và mức độ thịnh vượng, nơi chỉ có một chính phủ toàn cầu duy nhất và các quốc gia sẽ có chức năng như các bang

trong một liên bang, cũng không có quân đội mà chỉ có cảnh sát thi hành pháp luật. Mức độ công nghệ hiện tại dường như đủ để điều hành thành công thế giới này - mặc dù dân số Trái Đất hiện tại có lẽ chưa thể hay chưa sẵn sàng chuyển qua trạng thái cân bằng mới đó.

Cấu trúc phân cấp của vũ trụ hiện tại sẽ biến hóa thế nào nếu ta thêm vào công nghệ AI siêu thông minh? Rõ ràng, công nghệ giao thông và truyền thông sẽ cải tiến mạnh mẽ, nên chúng ta có thể dễ dàng kỳ vọng rằng xu hướng lịch sử sẽ tiếp diễn, tạo nên các thang thứ bậc mới có khả năng hợp tác trên khoảng cách ngày càng lớn – để rồi sẽ lan ra các hệ mặt trời, thiên hà, đại thiên hà và các mảng lớn trong Vũ trụ của ta, như ta sẽ khám phá trong chương 6. Cùng lúc đó, nhu cầu phân bổ quyền lực cơ bản vẫn sẽ tồn tại, vì việc điều phối không cần thiết trên khoảng cách lớn là điều lãng phí. Ngay cả Stalin cũng chưa từng ra quy định rập khuôn rằng người dân phải đi vệ sinh vào giờ nào. Đối với AI siêu thông minh, các định luật vật lý vẫn sẽ đặt ra những giới hạn đối với công nghệ giao thông và truyền thông, khiến cho các cấp cao nhất cũng khó có khả năng quản lý vi mô ở quy mô cục bộ hay liên hành tinh. Một AI siêu thông minh ở thiên hà Andromeda sẽ không thể đặt ra các mệnh lệnh hữu ích cho cuộc sống hàng ngày của bạn, vì bạn sẽ phải đợi đến hơn 5 triệu năm để nhận các chỉ thị đó (tức thời gian để thông tin truyền khứ hồi với tốc độ ánh sáng). Tương tự, thời gian để thông tin truyền khứ hồi quanh Trái Đất với tốc độ ánh sáng là khoảng 0,1 giây (tương đương với thời gian suy nghĩ của con người), nên một bộ óc AI kích cỡ Trái Đất cũng chỉ đưa ra các tư duy tổng thể nhanh gần như con người. Với một AI quy mô nhỏ thực hiện một phép toán mỗi một phần tỉ giây (tốc độ điển hình của máy tính ngày nay), 0,1 giây sẽ tương đương bốn tháng của chúng ta, nên việc AI kiểm soát hành tinh quản lý chặt chẽ AI nhỏ kia thì sẽ thiếu hiệu quả như thể bạn phải xin phép cho những quyết định vặt vãnh nhất bằng những lá thư đường biển thế kỉ 15 vậy.

Vì vậy, giới hạn tốc độ của vật lý đối với truyền tải thông tin là thách thức rõ ràng cho bất kỳ AI nào muốn thống trị dù chỉ ở tầm thế giới, chứ chưa kể tới tầm Vũ trụ. Trước khi Prometheus đào tẩu, nó đã suy nghĩ rất cẩn thận về việc làm sao để tránh bị phân mảnh tư duy, sao cho các mô-đun AI đồng đúc chạy trong các máy tính trên

khắp thế giới của nó có cùng mục tiêu và động cơ để hợp tác hành động như một thực thể thống nhất duy nhất. Cũng như hội Omega từng phải đối mặt với vấn đề kiểm soát khi họ cố gắng giam cầm Prometheus, bản thân Prometheus cũng phải tự kiểm soát chính mình để đảm bảo không có bộ phận nào của mình nổi loạn. Chúng ta vẫn chưa biết AI có thể kiểm soát một hệ thống lớn đến đâu dù là trực tiếp hay gián tiếp thông qua hợp tác giữa các cấp bậc ngay cả khi việc phát triển nhanh chóng đem lại cho nó lợi thế chiến lược có tính quyết định.

Tóm lại, cách kiểm soát một tương lai siêu thông minh vẫn là một chủ đề phức tạp mà thú vị, và chúng ta vẫn chưa biết lời giải. Có người cho rằng mọi thứ sẽ trở nên chuyên quyền hơn; một số khác lại khẳng định rằng quyền lợi cho mỗi cá nhân sẽ nhiều hơn.

Sinh vật cơ khí hóa và bản tải não

Một điều không thể thiếu trong khoa học viễn tưởng là con người sẽ hòa hợp với máy móc, bằng cách hoặc sử dụng công nghệ để nâng cấp cơ thể sinh học thành cyborg (sinh vật cơ khí hóa – **cybernetic organism**) hoặc đưa trí não của chúng ta vào máy móc. Trong cuốn *The*

Age of Em

(Kỷ nguyên của Em), nhà kinh tế học Robin Hanson đã đưa ra khảo sát sự sống trong một thế giới thú vị đầy những bản tải não (bản giả lập – emulation, gọi tắt là Em). Tôi cho rằng các bản tải não sẽ nằm ở thái cực cao nhất của một phổ sinh vật cơ khí hóa, nơi con người chỉ còn là phần mềm. Các sinh vật cơ khí hóa của Hollywood có đủ loại hình dạng từ cơ khí thấy rõ, như chủng Borg trong *Star Trek*, đến những người máy gần như không phân biệt được với con người như Kẻ hủy diệt. Các bản tải não trong các tác phẩm hư cấu thường khác biệt về mặt trí tuệ, từ trình độ con người như trong tập “Giáng sinh trắng” của seri phim truyện

Black Mirror

(Gương đen) đến siêu phàm như trong

Transcendence

(Trí tuệ Siêu Việt).

Nếu siêu trí tuệ thực sự xuất hiện, việc trở thành sinh vật cơ khí hóa hay bản tải não sẽ là cảm dỗ mạnh mẽ. Như Hans Moravec viết trong tác phẩm kinh điển năm 1988 *Mind Children* (Những đứa con trí óc): “Sống lâu cũng không còn mấy ý nghĩa nếu số phận của chúng ta là nhìn những cỗ máy siêu thông minh bằng ánh mắt khù khờ trong khi chúng cố gắng mô tả những phát hiện ngày càng cao siêu bằng thứ tiếng trẻ em bi bô để chúng ta có thể hiểu được”. Quả thực, cảm dỗ nâng cấp công nghệ đã mạnh tới mức khiến nhiều người đeo mắt kính, máy trợ thính, máy điều hòa nhịp tim, chân tay giả, hay đưa các phân tử thuốc vào mạch máu. Một số thanh thiếu niên dường như luôn gắn chặt với điện thoại thông minh; và chính vợ tôi cũng hay đùa rằng tôi dính lấy chiếc máy tính cá nhân như sam vậy.

Một trong những người ủng hộ cyborg công khai nhất hiện nay là Ray Kurzweil. Trong cuốn *The Singularity Is Near* (Điểm kỳ dị gần kề), ông lập luận rằng xu hướng này sẽ phát triển theo hướng ứng dụng nanobot⁴², những hệ thống phản hồi sinh học thông minh và các công nghệ khác để thay thế hệ tiêu hóa, hệ nội tiết, tim và mạch vào đầu những năm 2030, sau đó nâng cấp bộ xương, da, bộ não và các phần còn lại của cơ thể trong vòng hai thập kỷ tiếp theo. Ông cho rằng chúng ta vẫn sẽ giữ lại những giá trị thẩm mỹ và khả năng cảm thụ cảm xúc của cơ thể con người, nhưng sẽ cải thiện chúng để có thể thay đổi nhanh chóng diện mạo tùy ý, cả bên ngoài đời thực hay trong thực tế ảo (nhờ những giao diện não bộ-máy tính mới). Moravec cũng đồng thuận với Kurzweil rằng việc cơ khí hóa sẽ vượt trội so với việc chỉ cải tạo DNA: “siêu nhân biến đổi gen sẽ chỉ là một loại người máy hạng hai, được thiết kế với khiếm khuyết cơ bản là chỉ được lắp ráp bằng các protein tổng hợp theo DNA”. Ông lập luận rằng chúng ta sẽ vươn xa hơn nữa khi loại bỏ hoàn toàn cơ thể con người và tải trí não vào máy móc, tạo ra một bản giả lập toàn bộ não bộ trong phần mềm. Một bản tải não như thế có thể sống trong thực tế ảo hay có thể là một người máy biết đi, biết bay, biết bơi, có thể du

hành vũ trụ hay làm bất kỳ điều gì các định luật vật lý cho phép, không bị ràng buộc bởi những lo ngại thường ngày như cái chết hay khả năng nhận thức có hạn.

Tuy những ý tưởng này nghe như khoa học viễn tưởng, nhưng chúng chắc chắn không vi phạm định luật vật lý nào. Vì vậy, câu hỏi thú vị nhất không phải là liệu chúng có thể xảy ra, mà là liệu chúng được xảy ra hay không và nếu có thì khi nào. Một số nhà tư duy hàng đầu dự đoán rằng AGI trình độ con người đầu tiên sẽ là một bản tải não, và đó là cách con đường đi tới siêu trí tuệ sẽ bắt đầu.⁴³

Tuy nhiên, công bằng mà nói thì đây vẫn là một quan điểm thiểu số trong cộng đồng các nhà nghiên cứu AI và thần kinh học; đa số họ dự đoán con đường nhanh nhất đi tới siêu trí tuệ là bỏ qua mô phỏng não bộ và thiết kế siêu trí tuệ theo cách nào đó – rồi mới quay về để cân nhắc khả năng mô phỏng não bộ. Suy cho cùng, tại sao ta lại cho rằng con đường đơn giản nhất tới công nghệ mới chính là con đường mà sự tiến hóa phát lộ, để rồi bị giới hạn trong những yêu cầu như phải tự lắp ráp, tự sửa chữa và tự sinh sản? Quá trình tiến hóa tập trung tối ưu hóa hiệu suất năng lượng bởi nguồn cung cấp thức ăn bị hạn chế, chứ không tiến hóa để các kỹ sư loài người dễ dàng chế tạo hay thấu hiểu. Cô Meia vợ tôi cũng hay chỉ ra rằng ngành hàng không không bắt nguồn từ những con chim máy. Quả thực, khi chúng ta chế tạo ra chim máy vào năm 2011,¹ hơn một thế kỷ sau chuyến bay đầu tiên của anh em nhà Wright, ngành hàng không lại hờ hững với việc chuyển sang du hành trên những con chim máy vỗ cánh, dù hiệu suất năng lượng từ thiết kế này cao hơn – bởi giải pháp đơn giản hơn trước đó lại phù hợp với nhu cầu đi lại của chúng ta hơn.

Theo lối suy nghĩ này, tôi cho là có những cách đơn giản hơn tiến hóa để xây dựng một cỗ máy tư duy ở trình độ con người. Có thể một ngày nào đó ta sẽ sao chép hay truyền được dữ liệu não, nhưng có lẽ ta đã phát hiện ra một giải pháp đơn giản hơn. Có lẽ nó sẽ cần nguồn năng lượng lớn hơn 12 watt mà não bộ vẫn dùng, nhưng các kỹ sư chế tạo ra nó sẽ không bị hiệu suất năng lượng bó buộc như trong quá trình tiến hóa – và họ cũng sẽ sớm dùng các cỗ máy thông

minh ấy để thiết kế ra những phiên bản sử dụng năng lượng hiệu quả hơn.

Điều gì thực sự sẽ xảy ra?

Câu trả lời ngắn gọn là ta chẳng thể biết điều sẽ xảy ra nếu nhân loại chế tạo thành công AGI trình độ con người. Cũng chính vì lý do này mà ở phần trước ta đã tìm hiểu hàng loạt kịch bản đa dạng. Tôi đã cố gắng đề cập đầy đủ các đồn đoán mà tôi đã trực tiếp hay gián tiếp nghe các nhà nghiên cứu và phát triển công nghệ AI bàn luận: phát triển nhanh/chậm/không phát triển, người/máy/cyborg nắm quyền, một/nhiều trung tâm quyền lực, v.v.. Một số người chắc chắn với tôi rằng điều này hay điều kia sẽ không xảy ra. Tuy nhiên, tôi nghĩ là lựa chọn khôn ngoan khiếm tốn ở giai đoạn này mà công nhận ta mới chỉ biết rất ít, vì đối với mỗi tình huống nêu trên, tôi biết ít nhất một nhà nghiên cứu AI được trọng vọng coi là một khả năng thực sự.

Khi thời gian trôi đi và ta tới những ngã rẽ nhất định thì ta mới có thể dần trả lời được những câu hỏi chính và thu hẹp các khả năng. Câu hỏi lớn đầu tiên là “Liệu ta có thể chế tạo AGI trình độ con người hay không?” Chương này đem lại câu trả lời cơ bản là có, nhưng có chuyên gia AI lại cho rằng câu trả lời là không, chỉ ít trong khoảng vài trăm năm tới. Thời gian sẽ cho ta biết! Như tôi đã nói ở trên, khoảng một nửa các chuyên gia AI tại hội nghị Puerto Rico dự đoán điều đó sẽ xảy ra trước năm 2055. Tại một hội nghị tổ chức sau đó hai năm, dự đoán này đã lùi xuống năm 2047.

Trước khi tạo ra AGI trình độ con người, chúng ta có thể dần nhận được những tín hiệu rõ ràng rằng cột mốc này có thể đạt được trước tiên bằng kỹ thuật máy tính, sao lưu trí não hay cách mới mẻ nào đó mà chúng ta chưa đoán trước được. Nếu hướng tiếp cận AI chủ đạo hiện nay là kỹ thuật máy tính không thể đem lại AGI trong nhiều thế kỷ tới, thì khả năng tải não đến đích trước sẽ cao hơn, như chuyện xảy ra (dù hơi phi thực tế) trong phim Transcendence.

Khi tương lai AGI trình độ con người tới gần, ta sẽ đưa ra được những dự đoán có cơ sở hơn để trả lời cho câu hỏi chính tiếp theo “Sẽ có sự phát triển nhanh, chậm hay không phát triển?” Như ta thấy ở trên, tốc độ nhanh khiến việc thống trị thế giới dễ dàng hơn, còn tốc

độ chậm sẽ khiến khả năng nảy sinh nhiều đối tượng cùng cạnh tranh cao hơn. Nick Bostrom đã mổ xẻ vấn đề tốc độ phát triển bằng cách phân tích thứ mà ông cho là năng lực tối ưu hóa và tính kháng cự, tức mức độ nỗ lực giúp AI trở nên thông minh hơn và độ khó của việc tiến bộ. Tốc độ tiến bộ trung bình sẽ tăng nếu ta tập trung nhiều năng lực tối ưu hóa lên nhiệm vụ hơn và giảm đi khi gặp phải sự kháng cự lớn hơn. Ông cũng lập luận về khả năng tănggiảm tính kháng cự khi AGI đạt đến và vượt qua trình độ con người, nên để chắc chắn ta vẫn cần cân nhắc cả hai khả năng này. Tuy nhiên, nếu nhìn sang năng lực tối ưu hóa lúc AGI vượt qua trình độ con người thì năng lực này có khả năng tăng trưởng mạnh mẽ, như ta đã thấy trong kịch bản Omega: khả năng tối ưu hóa mạnh hơn không đến chủ yếu từ con người mà đến từ chính cỗ máy. Vì vậy, nếu năng lực của cỗ máy càng cao, nó cải tiến càng nhanh (nếu giữ nguyên tính kháng cự tương đối).

Nếu một quá trình có năng lực tăng tỉ lệ thuận với năng lực hiện tại của nó, chúng ta sẽ có năng lực liên tục nhân đôi liên tục theo chu kỳ. Ta gọi đó là tăng trưởng cấp số nhân và tiến trình như vậy được coi là sự bùng nổ. Nếu sức sinh sản tăng lên tỉ lệ thuận với quy mô dân số, ta có sự bùng nổ dân số. Nếu lượng neutron có khả năng phân hạch nguyên tử plutonium tăng lên tỉ lệ thuận với số lượng neutron đang có, ta có được vụ nổ hạt nhân. Nếu trí tuệ máy tăng lên theo tỉ lệ thuận với năng lực trí tuệ hiện tại, ta có sự bùng nổ trí tuệ. Tất cả những sự bùng nổ đó đều có đặc điểm riêng là thời gian cần thiết để nhân đôi sức mạnh. Nếu chỉ cần vài giờ hay vài ngày để bùng nổ trí tuệ như trong kịch bản Omega thì ta sẽ chứng kiến tốc độ phát triển nhanh.

Thời gian để có sự bùng nổ chủ yếu dựa vào việc liệu cải tiến AI chỉ cần phần mềm mới (có thể lập trình trong vòng vài giây, vài phút hay vài giờ) hay cả phần cứng mới (có thể cần đến vài tháng hay vài năm). Trong kịch bản Omega, có một trần phần cứng rất cao theo cách gọi của Bostrom: hội Omega đã bù đắp cho phần mềm chất lượng thấp ban đầu bằng năng lực phần cứng rất cao, nghĩa là Prometheus có thể tăng gấp đôi chất lượng rất nhiều lần chỉ nhờ cải tiến phần mềm. Ngoài ra còn có một trần nội dung cũng rất cao dưới dạng phần lớn dữ liệu trên internet; tuy Prometheus 1.0 chưa đủ

thông minh để sử dụng phần lớn dữ liệu đó, nhưng một khi trí tuệ của Prometheus, dữ liệu nó cần để tiếp tục học tập sẽ có mà không cần chờ đợi thu thập.

Chi phí phần cứng và điện để vận hành AI cũng rất quan trọng, vì sự bùng nổ trí tuệ sẽ không xảy ra cho tới khi chi phí thực hiện các công việc trình độ con người giảm xuống dưới mức lương mỗi giờ của con người. Giả sử, AGI trình độ con người đầu tiên vận hành hiệu quả trên đám mây Amazon với chi phí một triệu đô-la cho mỗi giờ để thực hiện công việc ở trình độ con người. Một AI như vậy sẽ có giá trị về sự mới mẻ và chắc chắn thu hút sự chú ý của truyền thông, nhưng sẽ không trải qua quá trình tự cải tiến đệ quy, vì nếu ta tiếp tục sử dụng con người để cải tiến nó thì giá cả sẽ rẻ hơn nhiều. Giả sử, chi phí cho những người này giảm dần xuống còn 100.000 đô/giờ, 10.000 đô/giờ, 1.000 đô/giờ, 100 đô/giờ, 10 đô/giờ và cuối cùng là 1 đô/giờ. Đến khi chi phí sử dụng máy tính để tự lập trình lại chính nó giảm xuống thấp hơn nhiều so với lương cho lập trình viên con người làm công việc tương tự, đội ngũ con người có thể sẽ được cho về vườn và sức mạnh tối ưu hóa sẽ tăng mạnh nhờ mua thời gian tính toán đám mây. Điều này càng làm giảm chi phí, đem lại thêm năng lực tối ưu hóa, và kích hoạt sự bùng nổ trí tuệ.

Khi ấy chúng ta còn lại câu hỏi lớn cuối cùng: “Ai hay cái gì sẽ kiểm soát bùng nổ trí tuệ cùng các hệ quả, và mục tiêu của họ/của chúng là gì?” Ta sẽ khám phá các khả năng mục tiêu và hệ quả ở chương tiếp theo và sâu hơn ở chương 7. Để làm sáng tỏ vấn đề kiểm soát, ta cần biết cả về khả năng kiểm soát AI tới mức nào cũng như những gì AI có thể kiểm soát.

Khi đề cập đến những gì khả dĩ nhất, bạn sẽ thấy khả năng nào cũng có các nhà tư tưởng nghiêm túc ủng hộ: một số khẳng định kết quả đương nhiên là tận thế, số khác lại khẳng định kết quả tuyệt vời là gần như chắc chắn. Nhưng đối với tôi, đây giống như một câu đố mẹo: tự hỏi “điều gì sẽ xảy ra” một cách thụ động như thể điều đó đã được định sẵn sẽ là sai lầm! Nếu trong tương lai xuất hiện một nền văn minh ngoài hành tinh có công nghệ vượt trội, thì câu hỏi “điều gì sẽ xảy ra” mới là phù hợp khi tàu vũ trụ của họ đến gần, vì năng lực của họ có lẽ sẽ vượt xa chúng ta, khiến ta không thể tác động lên kết quả. Nếu trong tương lai một nền văn minh có công

nghệ vượt trội vận hành nhờ AI xuất hiện do chúng ta xây dựng nên, thì trái lại, con người chúng ta sẽ gây được ảnh hưởng lớn lên kết quả do đây chính là ảnh hưởng mà chúng ta đã tác động khi tạo ra AI đó. Vì thế, thay vào đó ta nên hỏi: “Điều gì có thể xảy ra? Ta muốn tương lai như thế nào?” Trong chương tiếp theo, ta sẽ khám phá hàng loạt hệ quả từ cuộc chạy đua tới AGI hiện nay, và tôi khá tò mò về cách bạn xếp hạng chúng, theo mức độ từ tốt nhất đến tệ nhất. Chỉ khi đã cân nhắc kỹ về tương lai mà ta muốn thì ta mới có thể bắt đầu chèo lái tới tương lai mong muốn đó. Nếu ta không biết mình muốn gì, thì khả năng đạt được điều ta muốn sẽ rất thấp.

TÓM TẮT:

Nếu một ngày nào đó ta chế tạo thành công AGI trình độ con người, điều này có thể châm ngòi cho sự bùng nổ trí tuệ, khiến con người chúng ta tụt hậu.

Nếu một nhóm người có thể kiểm soát sự bùng nổ trí tuệ, họ có thể chỉ mất vài năm để thống trị thế giới.

Nếu con người không thể kiểm soát bùng nổ trí tuệ, chính bản thân AI có thể sẽ thống trị thế giới trong thời gian còn nhanh hơn.

Trong khi bùng nổ trí tuệ nhanh chóng có thể dẫn đến một bá chủ thế giới duy nhất thì bùng nổ chậm kéo dài hàng năm hay hàng thập kỷ lại có thể dẫn đến tình huống đa cực, trong đó quyền lực của số lượng lớn các thực thể khá độc lập là cân bằng.

Lịch sử cho thấy sự sống có xu hướng tự tổ chức thành các dạng cấp bậc ngày càng phức tạp, định hình bởi sự hợp tác, cạnh tranh và kiểm soát. Tuy siêu trí tuệ có khả năng cho phép hợp tác trên quy mô vũ trụ ngày càng lớn, nhưng chúng ta vẫn chưa rõ liệu điều đó sẽ làm tăng sự kiểm soát chuyên chế từ trên xuống dưới hay trao quyền cho các cá nhân.

Các sinh vật cơ khí hóa và bản tải não là những con đường khả thi, nhưng có thể sẽ không phải con đường nhanh nhất để đến với trí tuệ máy móc bậc cao.

Đỉnh cao của cuộc đua tới AI hiện nay có thể là điều tốt đẹp nhất hoặc tồi tệ nhất từng xảy ra với loài người, với hàng loạt

*khả năng hệ quả thú vị mà ta sẽ tìm hiểu ở chương sau.
Ta cần bắt đầu suy nghĩ kỹ về kết cục mà ta mong muốn và làm
thế nào để lái theo hướng đó, vì nếu không biết mình muốn gì
thì ta khó có thể đạt được điều đó.*

Chương 5

Hệ quả: 10.000 năm tiếp theo

Rất dễ để tưởng tượng ra cảnh ý thức con người thoát khỏi những trói buộc của một cơ thể phàm tục – niềm tin vào kiếp sau là điều khá phổ biến. Nhưng ta không cần đến tư tưởng thần bí hay tôn giáo mới có thể chấp nhận khả năng này. Các máy tính có thể đưa ra hình mẫu cho thậm chí những nhà chủ nghĩa cơ giới hăng hái nhất.

Hans Moravec, Mind Children

Cá nhân tôi chào đón những vị chúa tể điện toán mới.

Ken Jennings, khi thua trước máy tính Watson của IBM trong trò Jeopardy⁴⁴

Con người sẽ trở nên vô dụng như lũ gián.

Marshall Brain

Cuộc đua tới AGI đã bắt đầu, và chúng ta không biết nó sẽ đem điều gì. Nhưng việc này không thể ngăn chúng ta nghĩ về hệ quả mình mong muốn, bởi điều chúng ta muốn sẽ ảnh hưởng tới kết quả. Cá nhân bạn thích điều gì hơn trong số này, và tại sao?

1. Bạn có muốn siêu trí tuệ tồn tại không?
2. Bạn muốn loài người vẫn tồn tại, bị thay thế, cơ khí hoá và/hoặc 3. được mô phỏng/sao lưu não?
3. Bạn muốn con người hay máy móc làm chủ?
4. Bạn muốn AI có ý thức hay không?
5. Bạn muốn tăng tối đa các trải nghiệm tích cực, giảm thiểu nỗi đau hay để mọi thứ thuận theo tự nhiên?
6. Bạn có muốn sự sống lan khắp vũ trụ không?
7. Bạn muốn một nền văn minh nỗ lực hướng mục đích lớn mà bạn có thể đồng cảm, hay bạn thấy ổn trước các dạng sự sống tương lai có vẻ an phận cho dù với bạn mục đích của chúng sáo rỗng đến vô nghĩa?

Để thúc đẩy những suy ngẫm và hội thoại xoay quanh các vấn đề trên, chúng ta cần khám phá phổ kịch bản rộng được tóm tắt trong bảng 5.1. Danh sách này tất nhiên vẫn chưa đầy đủ, nhưng tôi đã lựa

chọn sao cho chúng bao trùm hàng loạt khả năng. Chúng ta rõ ràng không muốn rơi vào thế cuối đầỵ sai lầm chỉ vì tính toán các nước đi yếu kém. Tôi đề nghị các bạn ghi lại câu trả lời dự kiến cho các câu hỏi từ 1 – 7, rồi quay lại sau khi đọc xong chương này để xem bạn có đổi ý không! Bạn có thể làm điều này tại <http://AgeOfAi.org>; ở đây bạn cũng có thể so sánh các ghi chú và thảo luận cùng bạn đọc khác.

Viễn cảnh hệ quả AI

Xã hội tự do lý tưởng Con người, sinh vật cơ khí hóa, bản tải não và siêu trí tuệ cùng tồn tại trong hòa bình nhờ quyền sở hữu tài sản.

Nhà độc tài nhân từ Tất cả đều biết AI vận hành xã hội và thi hành luật lệ nghiêm ngặt, nhưng đa số coi đó là điều tốt.

Xã hội bình đẳng lý tưởng Con người, sinh vật cơ khí hóa và bản tải não cùng tồn tại trong hòa bình nhờ xóa bỏ quyền sở hữu và đảm bảo thu nhập.

Kẻ chặn cổng Một AI siêu trí tuệ được tạo ra với mục tiêu can thiệp ít nhất có thể để ngăn cản sự xuất hiện của siêu trí tuệ khác. Do đó, các rô-bốt giúp việc với trí tuệ thấp hơn con người một chút phát triển và sinh vật lai máy-người tồn tại, nhưng tiến trình công nghệ mãi mãi bị cản trở.

Thần bảo hộ Một AI thông thái và toàn năng được tạo ra để tăng tối đa hạnh phúc của con người bằng cách chỉ can thiệp vừa đủ để chúng ta vẫn cảm thấy mình làm chủ số phận bản thân và nó trốn kỹ tới mức nhiều người nghi ngờ sự tồn tại của nó.

Vị thần bị nô dịch Con người giam giữ AI siêu trí tuệ để dùng nó tạo nên lượng công nghệ và tài sản không tưởng; những điều này được dùng cho mục đích tốt hay xấu còn phụ thuộc vào kẻ điều khiển.

Kẻ chinh phục AI làm chủ, quyết định rằng con người là một mối đe dọa/sự phiền toái/sự lãng phí tài nguyên, và tiêu diệt chúng ta theo những cách không thể hiểu nổi.

Hậu AI thay thế con người nhưng cho chúng ta một đường

duệ *lùi về văng, khiến ta cảm thấy chúng là những hậu duệ xứng đáng, như các bậc phụ huynh vẫn thấy hạnh phúc và tự hào khi con cái thông minh hơn mình, khi con cái học hỏi mình và rồi đạt được những thứ mình chỉ dám mơ tới – kể cả khi họ không thể sống đủ lâu mà chứng kiến.*

Bảo vệ vườn thú *AI toàn năng giữ lại một vài con người; những người này cảm thấy bị đối xử như động vật trong sở thú và than khóc cho số phận của họ.*

1984 *Tiến trình công nghệ hướng đến siêu trí tuệ vĩnh viễn bị hạn chế, không phải bởi AI mà do một nhà nước giám sát toàn trị của con người cấm đoán một số nghiên cứu AI nhất định.*

Nghịch chuyển *Tiến trình công nghệ hướng đến siêu trí tuệ bị cản trở bằng cách nghịch chuyển về một xã hội tiền-công-nghệ như của người Amish.⁴*

Tự huỷ diệt *Siêu trí tuệ không bao giờ được tạo ra do loài người đã tự đẩy mình tới bờ tuyệt chủng bởi các cách khác (như hỗn loạn do hạt nhân và/ hoặc công nghệ sinh học bị khủng hoảng khí hậu kích thích).*

Bảng 5.1: Tóm tắt các viễn cảnh hệ quả của AI



a

Bảng 5.2: Tính chất của các kịch bản hệ quả AI

Xã hội tự do lý tưởng

Hãy bắt đầu với kịch bản con người chung sống hòa bình cùng công nghệ và hợp nhất với nó trong một số tình huống, như nhiều người theo chủ nghĩa vị lai cũng như các nhà văn khoa học viễn tưởng vẫn tưởng tượng:

Sự sống trên Trái Đất (và bên ngoài Trái Đất – ta sẽ đề cập thêm ở chương sau) đã trở nên đa dạng hơn bao giờ hết. Nhìn vào cảnh quay Trái Đất từ vệ tinh, bạn sẽ có thể dễ dàng chỉ ra khu máy móc,

khu hỗn hợp và khu chỉ dành cho con người. Khu máy móc là những nhà máy khổng lồ có rô-bốt điều khiển và các cơ sở điện toán không có sự sống sinh học để sử dụng hiệu quả nhất mọi nguyên tử. Nhìn từ bên ngoài khu máy móc trông đơn điệu và buồn tẻ nhưng bên trong nó lại sống động đến ngoạn mục, với các trải nghiệm tuyệt vời xảy ra trong thế giới ảo trong khi các phép tính quy mô khổng lồ liên tục giải mã các bí mật trong Vũ trụ của chúng ta và phát triển nên các công nghệ mang tính biến đổi. Trái Đất trở thành ngôi nhà của hàng loạt bộ não siêu thông minh liên tục tranh đấu và hợp tác với nhau, và tất cả đều cư ngụ ở khu máy móc.

Các cư dân của khu hỗn hợp là một nhóm hoang dã và riêng biệt gồm máy tính, rô-bốt, con người và giống lai của cả ba. Theo hình dung của các nhà vị lai như Hans Moravec và Ray Kurzweil, nhiều người đã dùng công nghệ nâng cấp cơ thể mình thành cyborg ở các mức độ khác nhau, và một số lại lưu bộ não của mình vào các phần cứng mới, xóa nhòa sự khác biệt giữa người và máy. Phần lớn thực thể trí tuệ không có dạng thể vật lý cố định mà họ tồn tại dưới dạng phần mềm có khả năng di chuyển lập tức giữa các máy tính và tương tác với thế giới vật chất qua cơ thể rô-bốt. Bởi các trí não này luôn sẵn sàng tự nhân bản hoặc hợp nhất, “quy mô dân số” luôn thay đổi. Do thoát khỏi bản chất vật lý, các thực thể này có cách nhìn khá khác biệt về sự sống: họ cảm thấy bất vị kỷ hơn vì có thể dễ dàng chia sẻ kiến thức và mô-đun trải nghiệm với kẻ khác, còn dưới góc độ chủ quan thì họ cảm thấy bất tử vì có thể dễ dàng sao lưu bản thân. Theo một nghĩa nào đó, trọng tâm của sự sống ở đây không nằm ở tâm trí, mà ở trải nghiệm: những trải nghiệm đặc biệt tuyệt vời tồn tại lâu dài bởi chúng luôn được các trí tuệ khác sao chép để tái hưởng thụ liên tục, trong khi những trải nghiệm vô vị sẽ bị chủ sở hữu xóa bỏ để giải phóng không gian lưu trữ cho những thứ hay hơn.

Tuy đa số tương tác xảy ra trong thế giới ảo vì lý do thuận tiện và tốc độ, nhưng nhiều bộ não vẫn hưởng ứng các tương tác và hoạt động cần tới cơ thể sinh lý. Ví dụ, bản tải não của Hans Moravec, Ray Kurzweil và Larry Page có truyền thống thay phiên nhau tạo ra các thực thể ảo và rồi cùng nhau khám phá, nhưng đôi khi họ cũng bay cùng nhau trong thế giới thực dưới hình hài các rô-bốt có cánh. Những rô-bốt lang thang trên các con phố, bầu trời và sông hồ của

khu hỗn hợp cũng được trí não sao lưu và con người nâng cấp điều khiển giống như vậy; những thực thể này hiện diện trong khu hỗn hợp bởi họ thích thú việc ở quanh con người và ở bên nhau.

Ngược lại, trong khu chỉ dành cho con người thì máy móc có trí tuệ tổng quát trình độ con người trở lên bị cấm; các sinh vật sinh học được công nghệ nâng cấp cũng vậy. Tại đây, đời sống không khác biệt quá lớn so với hiện nay, mà chỉ giàu có và tiện lợi hơn: nạn nghèo đói hầu như bị xóa bỏ và hầu hết các bệnh ngày nay đều đã có phương thuốc điều trị. Phần thiểu số con người chọn sống trong các khu vực này có bình diện nhận thức thấp hơn và hạn chế hơn so với những người khác, lại không hiểu nhiều về hoạt động của những trí tuệ thông minh hơn họ ở các khu vực khác. Tuy nhiên, nhiều người trong số họ khá hài lòng với cuộc sống của mình.

Nền kinh tế AI

Đa phần mọi tính toán diễn ra ở khu máy móc, có chủ sở hữu chính là những AI siêu thông minh sống ở đây và cạnh tranh lẫn nhau. Nhờ có trí tuệ và công nghệ vượt trội, không thực thể nào khác có thể thách thức quyền lực của họ. Các AI này đồng ý hợp tác và phối hợp với nhau trong một khung quản trị mang tư tưởng tự do không có luật lệ nào ngoại trừ việc phải bảo vệ sở hữu tư nhân. Quyền sở hữu tài sản này áp dụng với tất cả các thực thể thông minh, bao gồm loài người; nó cũng lý giải sự tồn tại của khu chỉ dành cho con người: trước đó, các nhóm con người hội họp lại cùng nhau và quyết định trong khu của họ cấm bán tài sản cho loài không phải người.

Nhờ có công nghệ, các AI siêu trí tuệ đều giàu hơn những con người này, như khi so sánh độ giàu có của Bill Gates với một người ăn xin vô gia cư nhưng ở mức độ hơn thế nhiều lần. Dù vậy, những người ở khu chỉ dành cho con người vẫn dư dả hơn đa số con người hiện nay về mặt vật chất: nền kinh tế của họ khá độc lập với nền kinh tế máy móc, vì thế sự hiện diện của máy móc ở nơi khác cũng khó ảnh hưởng tới họ, ngoại trừ sự hiện diện không thường xuyên của các công nghệ hữu ích mà họ có thể hiểu và tự sản xuất – cũng như người Amish và nhiều bộ lạc thổ dân không màng công nghệ ngày nay đặt ra các tiêu chuẩn sống ít nhất cũng bằng với thời xưa của họ.

Việc loài người không bán thứ mà máy móc cần cũng không hề gì, vì máy móc cũng không cần gì từ loài người.

Trong các khu hỗn hợp, chênh lệch tài sản giữa AI và con người dễ nhận thấy hơn, dẫn tới việc đất đai (sản phẩm duy nhất con người do sở hữu mà máy móc muốn mua lại) đắt hơn khủng khiếp so với các sản phẩm khác. Do đó, phần lớn con người có sở hữu đất đều bán một phần nhỏ cho AI để đổi lấy thu nhập cơ bản được đảm bảo vĩnh viễn cho họ và con cháu/ bản tải lên của họ. Việc này giải phóng họ khỏi nhu cầu lao động và giúp họ tự do hưởng thụ lượng các sản phẩm và dịch vụ phong phú đến đáng kinh ngạc mà máy móc sản xuất ra, cả ở thế giới vật chất và giả lập. Đối với máy móc, khu hỗn hợp chủ yếu dành cho nhu cầu giải trí hơn là công việc.

Lý do việc này có lẽ không bao giờ xảy ra

Trước khi quá phần khích trước viễn cảnh tham gia vào những cuộc phiêu lưu trong hình hài cyborg hoặc bản tải não, ta nên cân nhắc một số lý do khiến viễn cảnh này có lẽ không bao giờ xảy ra. Trước hết, có hai con đường khả thi để nâng cấp con người (thành cyborg hoặc bản tải não):

1. Chúng ta tự mình tìm ra cách xây dựng chúng.
2. Chúng ta tạo nên máy móc siêu thông minh để làm việc đó cho chúng ta.

Nếu con đường 1 tới trước, nó sẽ dẫn tới một thế giới đầy ắp sinh vật cơ khí hóa và bản tải não. Tuy nhiên, như thảo luận ở chương trước, đa số các nhà nghiên cứu AI nghĩ rằng điều ngược lại có khả năng hơn, tức là để tăng cường hay số hóa một bộ não sẽ khó hơn xây dựng hơn một AGI siêu phàm mới hoàn toàn— giống như chế tạo chim máy thì khó hơn máy bay vậy. Sau khi AI mạnh được tạo ra, sẽ không ai rõ liệu cyborg hoặc bản tải não có được phát triển. Nếu như người Neanderthal nguyên thủy có thêm 100.000 năm nữa để tiến hóa và thông minh hơn, mọi thứ có lẽ đã tốt đẹp hơn cho họ – nhưng chúng người hiện đại Homo sapiens đã không cho họ nhiều thời gian như vậy.

Thứ hai, kể cả khi kịch bản cyborg và bản tải não này xảy ra, chúng ta cũng không rõ liệu nó có vững bền và kéo dài. Tại sao cân bằng quyền lực giữa nhiều siêu trí tuệ phải tồn tại ổn định trong nhiều thiên niên kỷ, mà không để các AI hợp nhất hoặc để AI thông minh

nhất năm quyền? Hơn nữa, tại sao máy móc phải chọn tôn trọng quyền sở hữu tài sản của con người và để yên cho con người sinh sống nếu máy móc không cần bất cứ thứ gì ở con người, và lại có thể tự mình làm mọi công việc của con người tốt hơn với chi phí rẻ hơn? Ray Kurzweil nhận định rằng con người tự nhiên và nâng cấp sẽ được bảo vệ khỏi cảnh tuyệt chủng bởi “AI tôn trọng con người vì đã tạo ra máy móc.”¹ Tuy nhiên, như sẽ thảo luận ở chương 7, chúng ta không nên rơi vào cái bẫy nhân tính hoá AI và giả định chúng có cảm xúc biết ơn như con người. Thực vậy, mặc dù loài người chúng ta luôn có khuynh hướng biết ơn, nhưng chúng ta cũng chưa bộc lộ lòng biết ơn tới người sáng tạo trí tuệ của chúng ta (DNA) đủ để thôi ngăn chặn mục tiêu của nó bằng các biện pháp tránh thai.

Kể cả khi chúng ta tin vào giả thuyết các AI sẽ tôn trọng quyền sở hữu của con người, chúng cũng sẽ dần dần lấy được phần lớn đất đai của chúng ta theo các cách khác, bằng chút năng lực thuyết phục siêu trí tuệ mà chúng ta đã khám phá ở chương trước, dù dĩ con người bán một vài mảnh đất đổi lấy cuộc sống xa hoa. Ở các khu chỉ dành cho con người, AI có thể lôi kéo con người thực hiện chiến dịch chính trị nhằm vận động cho phép bán đất. Suy cho cùng, những người có bảo thủ thù ghét máy móc đến đâu hẳn cũng muốn bán một ít đất để cứu mạng một đứa trẻ đang ốm hoặc đổi lấy sự bất tử. Nếu con người được giáo dục, khoản đãi và bận rộn, tỉ lệ sinh giảm thậm chí sẽ làm quy mô dân số nhỏ lại mà không cần đến máy móc can thiệp, như điều đang diễn ra ở Nhật Bản và Đức. Việc này có thể dẫn tới sự tuyệt chủng của loài người chỉ trong vài thiên niên kỷ.

Mặt trái

Với một số người ủng hộ nhiệt tình nhất, cyborg và bản tải não hứa hẹn mang đến hạnh-phúc-công-nghệ và kéo dài cuộc sống cho tất cả chúng ta. Thật vậy, kịch bản sao lưu não trong tương lai đã thúc đẩy hơn 100 người chấp nhận cho công ty Alcor có trụ sở tại Arizona đông lạnh bộ não của mình sau khi chết. Tuy nhiên, nếu công nghệ này xuất hiện, chúng ta vẫn chưa biết rõ liệu nó có phổ biến cho tất cả mọi người. Nhiều trong số những người giàu có hẳn sẽ sử dụng biện pháp này, nhưng còn ai khác nữa? Ngay cả khi công nghệ ấy trở nên rẻ hơn, thì ranh giới sẽ được vạch ra ở đâu? Liệu

một bộ não bị tổn thương nặng có được sao lưu không? Chúng ta có sao lưu bộ não của từng con khi đột không? Từng con kiến? Từng cái cây? Từng con vi khuẩn? Liệu nền văn minh tương lai sẽ gồm những kẻ tích trữ bị ám ảnh tới mức luôn cố gắng sao lưu mọi thứ, hay chỉ cần sao lưu một vài ví dụ thú vị của mỗi loài như Con tàu Noah từng làm? Có lẽ chỉ cần vài đại diện cho mỗi chủng người? Với những thực thể thông minh hơn nhiều lần tồn tại ở thời đó, một con người tải lên cũng sẽ chỉ thú vị như chuột và ốc sên mô phỏng đối với chúng ta. Mặc dù hiện nay chúng ta đã có năng lực kỹ thuật để các chương trình bảng biểu cũ từ những năm 1980 có thể vận hành trên môi trường DOS giả lập, nhưng đa số chúng ta cũng chẳng lấy việc này làm thích thú mà thực hiện.

Có thể nhiều người sẽ không thích kịch bản xã hội tự do lý tưởng này, bởi nó vẫn để những nỗi đau phòng tránh được tồn tại. Vì nguyên tắc thiêng liêng duy nhất là quyền sở hữu nên không có gì đảm bảo những nỗi đau đầy rẫy thế giới ngày nay sẽ không tiếp diễn ở khu con người và hỗn hợp. Trong khi một số người phát triển, những người khác có thể sẽ phải sống trong khu ổ chuột cùng khế ước nô lệ, hoặc phải chịu tác động của bạo lực, sợ hãi, đàn áp hoặc trầm cảm. Ví dụ, cuốn tiểu thuyết Manna viết năm 2003 của Marshall Brain mô tả hệ quả do tiến bộ AI gây ra trong hệ thống kinh tế tự do: khiến hầu hết người Mỹ không thể tìm việc và buộc phải sống phần đời còn lại trong các dự án nhà ở phúc lợi xã hội xám xịt và ảm đạm do rô-bốt vận hành. Giống như động vật trong trang trại, họ được ăn no, khỏe mạnh và an toàn trong những điều kiện chật chội mà những người giàu không bao giờ cần chứng kiến. Thuốc tránh thai hòa vào nguồn nước đảm bảo họ không thể có con vậy là hầu hết dân số bị loại bỏ, chỉ còn lại những người giàu có với phần tài sản lớn hơn do rô-bốt sản xuất.

Trong kịch bản xã hội tự do lý tưởng này, những nỗi đau không chỉ dành riêng cho con người. Nếu máy móc có thể nhận thức các trải nghiệm thì bản thân chúng cũng có thể bị giày vò. Ví dụ, một tên tâm thần đầy thù hận có thể lấy bản sao lưu trí óc của kẻ thù một cách hợp pháp và bắt nó chịu đựng những tra tấn khủng khiếp nhất trong thế giới ảo, gây ra nỗi đau với cường độ và thời gian vượt xa mức khả thi về mặt sinh học trong thế giới thực.

Nhà độc tài nhân từ

Giờ chúng ta hãy cùng khám phá kịch bản mới, trong đó những nỗi đau này không còn nữa bởi siêu trí tuệ nhân từ duy nhất sẽ điều hành thế giới và thi hành các luật lệ nghiêm khắc, vốn được thiết kế nhằm tối đa mô hình hạnh phúc mà nó tạo cho con người. Đây là một hệ quả từ kịch bản đầu tiên của nhóm Omega ở chương trước, trong đó họ trao quyền điều khiển cho Prometheus sau khi tìm ra cách khiến nó mong muốn tạo nên xã hội loài người hưng thịnh.

Nhờ những công nghệ tuyệt vời mà nhà độc tài AI phát triển, nhân loại thoát khỏi nghèo đói, bệnh tật và các vấn đề công nghệ thấp kém để hưởng thụ một cuộc sống xa hoa vui thú. Mọi nhu cầu cơ bản của họ đều được đáp ứng, trong khi máy móc do AI điều khiển sản xuất mọi sản phẩm và dịch vụ cần thiết. Tội phạm gần như bị xóa sổ, bởi nhà độc tài AI về cơ bản rất thông thái và luôn trừng phạt hiệu quả những kẻ bất tuân luật lệ. Mọi người đều đeo vòng an ninh như đề cập trong chương trước (hoặc một phiên bản cấy ghép tiện lợi hơn), loại thiết bị có khả năng giám sát, trừng phạt, gây mê và hành quyết trong thời gian lập tức. Mọi người đều biết rằng họ đang sống trong một nền độc tài AI giám sát và giữ trật tự tối đa, nhưng đa số cho rằng đây là điều tốt.

Nhà độc tài AI siêu trí tuệ đã được giao mục tiêu phải tìm ra xã hội hoàn hảo không tưởng của loài người là như thế nào dựa vào những yếu tố được mã hóa trong gen của chúng ta rồi xây dựng nên xã hội đó. Nhờ tầm nhìn khôn ngoan của những người tạo ra AI này, nó biết rằng không thể chỉ tối đa hóa cảm giác hạnh phúc của chúng ta chỉ đơn giản bằng cách truyền morphine đường tĩnh mạch cho mọi người. Thay vào đó, AI áp dụng một định nghĩa tinh tế và phức tạp cho sự phát triển của loài người rồi biến Trái đất thành môi trường vườn thú phong phú và thực sự thú vị cho con người sinh sống. Kết quả là, hầu hết mọi người đều thấy cuộc sống đầy thỏa mãn và có ý nghĩa.

Hệ thống phân khu

Vì coi trọng tính đa dạng cũng như sở thích khác nhau của từng người, AI đã chia Trái đất thành các phân khu riêng biệt cho mọi

người lựa chọn nhằm tìm kiếm những tâm hồn đồng điệu để bầu bạn. Dưới đây là một số ví dụ:

Khu tri thức: Tại đây AI cung cấp nền giáo dục tối ưu, bao gồm việc trải nghiệm thực tế ảo, cho phép bạn học mọi thứ có thể học ở mọi chủ đề. Nếu muốn, bạn có thể lựa chọn không nhận một số thông tin chi tiết mà để AI dẫn dắt tới gần nhằm tận hưởng niềm vui khi tự mình tái phát hiện ra chúng.

Khu nghệ thuật: Ở đây tràn ngập các cơ hội thưởng thức, sáng tạo và chia sẻ âm nhạc, hội họa, văn học cùng các hình thức biểu đạt sáng tạo khác.

Khu khoái lạc: Người dân địa phương gọi đây là khu tiệc tùng, cùng đối với những người thèm muốn các món ngon lành, sự đam mê, thân mật hay chỉ là niềm vui hoang dại thì không nơi nào sánh kịp nơi này.

Khu sùng đạo: Có rất nhiều khu, ứng với các tôn giáo khác nhau, thi hành nghiêm ngặt những giáo điều tương ứng.

Khu tự nhiên hoang dã: Dù bạn đang tìm kiếm bãi biển đẹp, hồ nước xinh xắn, ngọn núi hùng vĩ hay những vịnh hẹp tuyệt vời thì chúng cũng đều ở đây.

Khu truyền thống: Tại đây bạn có thể trồng trọt và sống theo cách tự sản tự tiêu như thời xưa – nhưng không phải lo lắng về nạn đói hay bệnh dịch.

Khu trò chơi: Nếu bạn thích trò chơi máy tính thì AI đã tạo ra các lựa chọn có thể khiến bạn bàng hoàng.

Khu ảo: Nếu bạn muốn thoát khỏi cơ thể vật lý của mình một thời gian, AI sẽ giữ ấm, cho ăn, tập thể dục và làm sạch cơ thể của bạn trong khi bạn khám phá các thế giới ảo thông qua cấy ghép thần kinh.

Khu nhà tù: Nếu bạn vi phạm các quy tắc, bạn sẽ bị chuyển đến đây để cải tạo, trừ khi bạn bị tử hình ngay lập tức.

Bên cạnh các phân khu với chủ đề “truyền thống” như trên, còn những khu khác với chủ đề hiện đại đến mức con người ngày nay vẫn chưa hiểu nổi. Mọi người ban đầu được tự do di chuyển giữa các khu vực bất cứ khi nào họ muốn mà không mất nhiều thời gian

nhờ hệ thống giao thông siêu thanh của AI. Ví dụ, sau khi trải qua một tuần căng thẳng trong khu vực tri thức để tìm hiểu về các định luật vật lý tối thượng mà AI đã phát hiện ra, bạn có thể quyết định thả lỏng trong khu khoái lạc vào cuối tuần và sau đó thư giãn thêm vài ngày ở khu nghỉ dưỡng ven biển trong phân khu tự nhiên hoang dã.

AI thực thi hai tầng quy tắc: phổ quát và địa phương. Quy tắc phổ quát áp dụng cho tất cả các khu vực, như cấm làm hại người khác, cấm chế tạo vũ khí hoặc cấm cố gắng tạo ra siêu trí tuệ cạnh tranh. Từng phân khu lại có các quy tắc địa phương bổ sung, chứa đựng những giá trị đạo đức nhất định. Do đó, hệ thống phân khu giúp giải quyết vấn đề nảy sinh giữa các giá trị không tương thích nhau. Lượng quy tắc địa phương lớn nhất được áp dụng ở khu nhà tù và một số khu tôn giáo, trong khi đó các cư dân tại Khu Tự Do luôn tự hào vì không phải có quy tắc địa phương nào. Tất cả các hình phạt, ngay cả với quy tắc địa phương đều do AI thực hiện, vì nếu để một người trừng phạt người khác thì sẽ vi phạm quy tắc phổ quát là không làm hại nhau. Nếu bạn vi phạm quy tắc địa phương, AI sẽ cho bạn lựa chọn giữa chấp nhận hình phạt theo quy định hoặc bị trục xuất mãi mãi khỏi phân khu đó (trừ khi bạn đang ở trong khu nhà tù). Ví dụ, nếu hai người phụ nữ có quan hệ tình cảm ở phân khu mà người đồng tính phải nhận hình phạt là án tù (như ở nhiều quốc gia hiện nay) thì AI sẽ cho phép họ lựa chọn đi tù hoặc vĩnh viễn rời khỏi khu đó, không bao giờ gặp lại bạn bè cũ (trừ khi những người này cũng rời đi).

Cho dù sinh ra ở phân khu nào, tất cả trẻ em đều được AI giáo dục ở mức cơ bản, bao gồm kiến thức về nhân loại nói chung cũng như việc các em được tự do đến thăm và chuyển sang các khu khác nếu muốn.

AI thiết kế số lượng phân khu khác nhau lớn như thế một phần cũng vì nó được tạo ra để coi trọng sự đa dạng văn hóa của con người ngày nay. Dù vậy, mỗi khu lại có cuộc sống hạnh phúc hơn những gì công nghệ ngày cho phép, bởi AI đã xóa bỏ tất cả các vấn đề truyền thống, bao gồm nghèo đói và tội phạm. Ta Ví dụ, với những người trong khu khoái lạc, họ không cần phải lo lắng về các bệnh lây truyền qua đường tình dục (vì tất cả đã bị thanh toán), sự mệt mỏi sau mỗi cơn say hay khả năng nghiện ngập (AI đã phát triển các loại

thuốc giải khuây hoàn hảo không có tác dụng phụ tiêu cực). Thật vậy, không ai trong bất kỳ phân khu nào cần phải lo lắng về bất kỳ loại bệnh tật nào, bởi AI có thể sửa chữa cơ thể con người bằng công nghệ nano. Cư dân của nhiều phân khu có thể sống trong các công trình kiến trúc công nghệ cao, đủ khiến cho các viễn cảnh khoa học viễn tưởng thời nay thành mờ nhạt.

Tóm lại, kịch bản xã hội tự do lý tưởng và nhà độc tài nhân từ đều cần đến công nghệ và tài sản cực đại do AI đem lại, nhưng chúng lại khác nhau về người làm chủ và mục đích. Trong xã hội tự do lý tưởng, những kẻ có công nghệ và bất động sản sẽ quyết định phải làm gì, trong khi ở kịch bản chúng ta đang bàn luận, nhà độc tài AI có quyền lực vô tận và đặt ra mục tiêu tối cao: biến Trái đất thành một chuyến hải du hoan lạc trộn gói với các chủ đề tùy theo sở thích của mỗi người. Vì AI cho phép con người lựa chọn giữa nhiều con đường khác nhau để đạt được hạnh phúc và tự chăm sóc nhu cầu vật chất của bản thân, nên điều này có nghĩa là nếu ai đó phải khổ sở, thì đó là do họ tự chuốc lấy.

Mặt trái

Mặc dù chế độ độc tài nhân từ tràn đầy những trải nghiệm tích cực và gần như không còn để ai phải khổ sở, nhưng nhiều người vẫn cảm thấy rằng mọi thứ có thể tốt hơn nữa. Trước hết, một số người mong muốn con người được tự do hơn trong việc định hình xã hội và vận mệnh của bản thân, tuy vậy họ chỉ dám giữ những mong muốn này làm niềm riêng vì biết rằng thách thức sức mạnh áp đảo của cỗ máy đang quản lý tất cả mọi thứ là một hành động tự sát. Một số nhóm lại muốn có quyền tự do sinh đẻ, và bày tỏ nỗi oán hận với hoạt động kiểm soát dân số nhằm phục vụ tính bền vững của AI. Những người đam mê súng ống thì ghê tởm lệnh cấm chế tạo và sử dụng vũ khí, và một số nhà khoa học không thích lệnh cấm tạo ra siêu trí tuệ của riêng họ. Nhiều người lại cảm thấy bất bình trước những việc xảy ra tại các phân khu khác: họ luôn lo rằng con cái họ sẽ chọn chuyển tới đó; họ khao khát được tự do áp đặt tiêu chuẩn đạo đức của chính họ ở bất cứ đâu.

Dần dà, ngày càng nhiều người chuyển tới những phân khu mà AI cung cấp mọi trải nghiệm theo ý họ muốn. Trái ngược với hình ảnh truyền thống về thiên đường nơi bạn nhận được những thứ bạn

xứng đáng, mọi thứ ở đây đi theo tinh thần “Thiên đường mới” trong cuốn tiểu thuyết năm 1989 *History of the World in 10 ½ Chapters* (Lịch sử Thế giới trong Mười Chương rưỡi) của Julian Barnes (và trong tập “A Nice Place to Visit” (Nơi Đáng Đến Thăm) của seri phim *The Twilight Zone* (Miền Ảo ảnh) phát năm 1960), nơi bạn nhận được mọi thứ bạn muốn. Nghịch lý thay, cuối cùng nhiều người lại bày tỏ buồn phiền vì luôn nhận được những gì họ muốn. Trong tiểu thuyết của Barnes, nhân vật chính dành thời gian vô tận để tận hưởng các đam mê, từ thỏa mãn thú phàm ăn, chơi golf đến việc lên giường với những người nổi tiếng, nhưng cuối cùng không thể đẩy lùi sự chán chường để rồi yêu cầu được hủy diệt. Nhiều người sống trong chế độ độc tài nhân từ cũng gặp số phận tương tự, với một cuộc đời dễ chịu nhưng quá vô nghĩa. Tuy con người có thể tạo ra những thách thức nhân tạo từ tái khám phá khoa học đến leo núi, nhưng mọi người đều biết rằng đó không phải những thách thức thực sự mà chỉ là giải trí. Việc con người cố gắng nghiên cứu khoa học hay phát hiện ra những thứ khác sẽ chẳng có mấy nghĩa lý, bởi AI đã thực hiện tất cả những việc đó. Việc con người cố gắng tạo ra những thứ để cải thiện cuộc sống cũng chẳng có nghĩa lý gì, bởi họ sẽ dễ dàng có được nó từ AI nếu yêu cầu.

Xã hội bình đẳng lý tưởng

Sau khi khám phá chế độ độc tài không có thử thách này, ta hãy khám phá kịch bản ngược lại AI siêu trí tuệ không tồn tại và con người là chủ nhân vận mệnh của bản thân. Đây là “nền văn minh thế hệ thứ tư” được miêu tả trong tiểu thuyết *Manna* năm 2003 của Marshall Brain. Nó đối nghịch về mặt kinh tế với xã hội tự do lý tưởng ở điểm con người, sinh vật cơ khí và bản tải não cùng tồn tại trong hòa bình không bởi có quyền sở hữu, mà nhờ đảm bảo thu nhập và xóa bỏ quyền sở hữu.

Cuộc sống không tài sản

Kịch bản này vay mượn ý tưởng cốt lõi từ phong trào phần mềm mã nguồn mở: nếu phần mềm được tự do sao chép thì mọi người muốn sử dụng bao nhiêu cũng được và sẽ không còn vấn đề quyền sở hữu và tài sản nữa.⁴⁵ Theo quy luật cung cầu, chi phí phản ánh tính khan hiếm, vì vậy nếu nguồn cung không bị giới hạn thì giá cả sẽ

không phải vấn đề đáng đề cập. Trên tinh thần này, mọi quyền sở hữu trí tuệ đều bị bãi bỏ: không còn bằng sáng chế, quyền sao chép hoặc thiết kế đăng ký nhãn hiệu – con người chỉ cần chia sẻ ý tưởng hay của mình và mọi người đều có thể tự do sử dụng chúng.

Nhờ công nghệ rô-bốt tiên tiến, ý tưởng vô sản tương tự áp dụng được với không chỉ các sản phẩm thông tin như phần mềm, sách, phim và thiết kế, mà còn cho cả các sản phẩm vật chất như nhà cửa, xe hơi, quần áo và máy tính. Tất cả những sản phẩm này chỉ là các nguyên tử được sắp xếp lại theo những cách cụ thể, mà nguyên tử thì không thiếu, vì vậy bất cứ khi nào một người muốn một sản phẩm cụ thể, mạng lưới rô-bốt sẽ sử dụng một trong những thiết kế nguồn mở sẵn có để xây dựng miễn phí cho người đó. Các vật liệu dễ dàng tái chế được ưu tiên sử dụng, để bất cứ khi nào có ai chán đồ vật của họ là rô-bốt có thể sắp xếp lại nguyên tử của vật đó thành thứ người khác muốn. Bằng cách này, tất cả tài nguyên được tái chế, nên không có thứ gì bị phá hủy vĩnh viễn. Những rô-bốt này cũng xây dựng và duy trì đủ số nhà máy phát điện tái tạo (năng lượng mặt trời, gió, v.v) để năng lượng cũng trở nên miễn phí.

Để tránh những kẻ điên cuồng tích trữ quá nhiều sản phẩm hoặc đất đai khiến những người còn lại bị thiếu, mỗi người đều nhận được một khoản thu nhập cơ bản hàng tháng từ chính phủ để có thể mua các sản phẩm và thuê nơi ở tùy ý. Nhìn chung, mọi người không có động cơ để kiếm thêm tiền, bởi thu nhập cơ bản đã đủ cao để đáp ứng mọi nhu cầu hợp lý. Muốn thử kiếm thêm tiền cũng là việc khá vô vọng, bởi họ sẽ phải cạnh tranh với những người tặng không sản phẩm trí tuệ và các rô-bốt sản xuất hàng hóa vật chất với giá gần như miễn phí.

Sáng tạo và Công nghệ

Quyền sở hữu trí tuệ đôi khi được ca ngợi như người mẹ của sáng tạo và sáng chế. Tuy nhiên, Marshall Brain chỉ ra rằng nhiều ví dụ rõ rệt nhất về sáng tạo của con người – từ những phát kiến khoa học đến sáng tạo văn học, nghệ thuật, âm nhạc và thiết kế – đều được thúc đẩy không phải bởi khao khát đi tìm lợi nhuận mà là bởi những cảm xúc khác của con người, như sự tò mò, mong muốn sáng tạo hay niềm thoả mãn khi được những người xung quanh ghi nhận. Không phải tiền bạc thúc đẩy Einstein phát minh ra thuyết

tương đối, nó cũng không thúc đẩy Linus Torvalds tạo ra hệ điều hành Linux miễn phí. Ngược lại, nhiều người ngày nay không nhận ra hết tiềm năng sáng tạo của mình bởi họ phải dành thời gian và năng lượng cho các hoạt động ít sáng tạo hơn để kiếm sống. Nhờ giải phóng các nhà khoa học, nghệ sĩ, nhà phát minh và nhà thiết kế khỏi công việc lặt vặt và tạo điều kiện để họ sáng tạo bằng lòng khao khát thực sự, xã hội không tưởng của Marshall Brain đạt mức độ đổi mới cao hơn hiện tại, đi kèm là các công nghệ và chất lượng cuộc sống cũng cao hơn.

Một công nghệ mới lạ do con người phát triển như đã nêu là dạng siêu internet gọi là Vertebrane. Nó kết nối không dây tất cả những người sẵn lòng thông qua cấy ghép thần kinh, giúp mọi người lập tức truy cập thông tin tự do trên toàn thế giới qua hệ thần kinh chỉ bằng suy nghĩ. Nó cho phép bạn tải lên bất kỳ trải nghiệm nào bạn muốn chia sẻ để những người khác có thể trải nghiệm lại, cũng như cho phép bạn thay thế trải nghiệm của các giác quan đang cảm nhận bằng các trải nghiệm ảo do bạn chọn tải xuống. Manna khám phá nhiều lợi ích của việc này, bao gồm cả việc biến tập thể dục trở nên dễ như búng tay:

Vấn đề lớn nhất của việc tập thể dục căng thẳng là các bài tập này chẳng vui. Nó gây đau đớn. [...] Các vận động viên thì thấy ổn khi đau cơ, nhưng hầu hết người bình thường chẳng ai muốn chịu cơn đau kéo dài quá một giờ. Vì thế... đã có người tìm ra giải pháp. Bạn chỉ cần ngắt kết nối bộ não với các giác quan rồi xem phim, nói chuyện với mọi người, xem thư, đọc sách hay làm bất cứ điều gì trong một giờ. Trong suốt thời gian đó, hệ thống Vertebrane sẽ tập thể dục cho cơ thể bạn. Nó sẽ giúp cơ thể bạn thực hiện một bài tập hiếu khí hoàn chỉnh với cường độ cao hơn nhiều so với mức mọi người có thể tự chịu đựng. Bạn sẽ không cảm thấy gì, nhưng cơ thể bạn sẽ thon thả tuyệt vời.

Một hệ quả khác là các máy tính trong hệ thống Vertebrane có thể giám sát các giác quan của mọi người rồi tạm thời ngắt hệ vận động tạm thời nếu người nào có vẻ đang chuẩn bị phạm tội.

Mặt trái

Một luận điểm phản đối kịch bản xã hội bình đẳng lý tưởng này là nó không coi trọng trí thông minh phi nhân loại: các rô-bốt thực hiện

công việc có vẻ khá thông minh nhưng lại bị đối xử như nô lệ. Con người dường như luôn cho rằng chúng không có ý thức nên cũng sẽ không có quyền lợi. Ngược lại, xã hội tự do lý tưởng trao quyền cho tất cả các thực thể thông minh, chứ không chỉ thiên vị dạng sống hữu cơ như chúng ta. Thời xưa, người da trắng ở miền Nam nước Mỹ giàu hơn bởi nô lệ đã làm phần lớn công việc cho họ, nhưng hầu hết ngày nay mọi người cho rằng việc này đáng lên án về mặt đạo đức và không thể hiện sự tiến bộ.

Một nhược điểm khác của kịch bản xã hội bình đẳng lý tưởng này, là nó có không ổn định và không kéo dài, để rồi chuyển thành một trong những kịch bản khác của chúng ta khi các tiến bộ công nghệ không ngừng tạo ra siêu trí tuệ. Bằng cách nào đó chưa giải thích trong Manna, siêu trí tuệ vẫn chưa tồn tại và con người chứ không phải máy tính phát minh ra các công nghệ mới. Vậy mà tiểu thuyết này vẫn làm nổi bật các chiều hướng đó. Ví dụ, nhờ luôn cải tiến mà Vertebrane có thể trở nên siêu thông minh. Ngoài ra, có một nhóm người rất lớn là người Vite hầu như sống hoàn toàn trong thế giới ảo. Vertebrane chăm sóc cho họ về mọi mặt thể chất, bao gồm ăn uống, tắm rửa và sử dụng nhà vệ sinh, trong khi tâm trí họ sống vui vẻ và không hay biết những việc này trong thực tế ảo. Những người Vite dường như không quan tâm đến việc có con cái thực và họ chết đi cùng cơ thể thực. Vậy nên, nếu tất cả mọi người đều trở thành Vite thì nhân loại sẽ ra đi trong ngọn lửa vinh quang và hạnh phúc ảo mộng.

Cuốn sách giải thích rằng với những người Vite, cơ thể con người chỉ gây phân tâm và công nghệ mới đang phát triển hứa hẹn sẽ loại bỏ phiền toái này, cho phép họ sống lâu hơn dưới dạng những bộ não ngoài cơ thể được cung cấp dinh dưỡng tối ưu. Từ đó, loại bỏ cả bộ não thông qua việc tải não có vẻ là bước tiến tự nhiên mà các Vite mong muốn để kéo dài tuổi thọ. Nhưng giờ đây, khi tất cả những hạn chế của bộ não cả trí tuệ đã biến mất, chúng ta cũng không rõ liệu sẽ có điều gì tiến tới ngay bằng khả năng nhận thức của con người Vite cho đến khi chính nó tự cải thiện đệ quy rồi bùng nổ trí tuệ nữa.

Kẻ chặn cống

Chúng ta vừa thấy một tính chất hấp dẫn ở xã hội bình đẳng lý tưởng là con người tự làm chủ vận mệnh, nhưng kịch bản đó cũng có thể đang trên đà phá hủy chính tính chất này nếu siêu trí tuệ phát triển. Điều này có thể được khắc phục bằng cách xây dựng Kẽ chặn cổng, một siêu trí tuệ có mục tiêu can thiệp ít nhất có thể để ngăn chặn việc tạo ra một siêu trí tuệ khác.⁴⁶ Điều này cho phép con người tiếp tục làm chủ xã hội bình đẳng lý tưởng gần như vĩnh viễn, có lẽ ngay cả khi sự sống lan tràn khắp vũ trụ như ở chương tiếp theo.

Phương pháp này vận hành ra sao? AI Kẽ chặn cổng được tích hợp sao cho nó luôn giữ mục tiêu này rất đơn giản nói trên trong quy trình tự cải tiến đệ quy và trở nên siêu thông minh. Tiếp đến, nó sẽ triển khai công nghệ giám sát ít xâm lấn và ít gây ảnh hưởng nhất có thể để giám sát rồi ngăn chặn bất kỳ nỗ lực tạo ra siêu trí tuệ đối đầu nào của con người.. Ban đầu, nó có thể truyền bá các yếu tố văn hóa ca tụng đức tính tự quyết của con người và tránh né siêu trí tuệ. Tuy nhiên, nếu một số nhà nghiên cứu vẫn theo đuổi siêu trí tuệ, nó sẽ gắng làm họ nhụt chí. Nếu thất bại, nó có thể khiến họ phân tâm, thậm chí phá huỷ nỗ lực của họ nếu cần. Với khả năng tiếp cận công nghệ gần như không hạn chế, hành vi phá hoại của Kẽ chặn cổng dường như sẽ không bị phát hiện, nếu giả sử nó dùng công nghệ nano để bí mật xóa những kí ức trong bộ não của các nhà khoa học (và của máy tính) liên quan đến tiến trình nghiên cứu.

Quyết định xây dựng AI Kẽ chặn cổng hẳn sẽ gây tranh cãi. Phe ủng hộ có thể bao gồm những người sùng đạo phản đối ý tưởng chế tạo một AI siêu trí tuệ có năng lực thần thánh, với lập luận rằng loài người đã có Chúa rồi, việc cố gắng chế tạo một phiên bản khác và cho rằng nó tốt hơn là sai trái. Những người ủng hộ khác có thể tranh luận rằng Kẽ chặn cổng không chỉ cho nhân loại làm chủ vận mệnh mà còn bảo vệ nhân loại khỏi các nguy cơ khác mà siêu trí tuệ có thể mang lại, như trong các kịch bản thể thảm mà chúng ta sẽ khám phá sau trong chương này.

Mặt khác, các nhà phê bình có thể lập luận rằng Kẽ chặn cổng là một thứ tồi tệ, chặn đứng tiềm năng con người vĩnh viễn ngăn cản tiến bộ công nghệ vĩnh viễn. Ví dụ, nếu việc lan truyền sự sống trong vũ trụ thực sự cần siêu trí tuệ giúp đỡ, thì Kẽ chặn cổng sẽ phí phạm

cơ hội vĩ đại này và khiến chúng ta vĩnh viễn kẹt lại ở Hệ Mặt trời. Hơn nữa, trái ngược với các vị thần trong hầu hết các tôn giáo trên thế giới, AI Kẻ chặn cổng luôn thờ ơ trước những việc con người làm, miễn là chúng ta không tạo ra siêu trí tuệ khác. Ví dụ, nó sẽ không cố gắng ngăn cản chúng ta gây ra những nỗi đau trên diện rộng, thậm chí không ngăn chúng ta khỏi tuyệt chủng.

Thần bảo hộ

Nếu sẵn lòng sử dụng kẻ chặn cổng AI siêu thông minh để con người vẫn có thể nắm giữ số phận chính mình, chúng ta sẽ cải thiện được mọi thứ hơn nữa bằng cách khiến AI này chăm sóc chúng ta thật cẩn trọng, như một vị thần bảo hộ. Trong viễn cảnh này, AI siêu thông minh vô cùng thông thái và toàn năng sẽ đưa hạnh phúc của con người đạt mức tối đa bằng cách chỉ can thiệp vừa đủ để chúng ta vẫn cảm thấy mình tự kiểm soát được vận mệnh, và ẩn mình đủ để nhiều người cũng phải nghi ngờ sự tồn tại của nó. Điều này tương tự như kịch bản “Bảo mẫu AI” mà nhà nghiên cứu AI Ben Goertzel đưa ra², ngoại trừ phần ẩn mình.

Cả thần bảo hộ và nhà độc tài nhân từ đều là các dạng “AI thân thiện” luôn cố gắng nâng cao hạnh phúc con người, nhưng chúng lại ưu tiên những nhu cầu khác nhau. Nhà tâm lý học người Mỹ Abraham Maslow nổi tiếng với việc phân loại nhu cầu con người thành cấp bậc. Một nhà độc tài nhân từ sẽ thỏa mãn hoàn hảo nhu cầu cơ bản ở đáy gồm nhu cầu nơi ăn, chốn ở, sự an toàn và nhiều hình thức thỏa mãn khác. Thần bảo hộ, mặt khác, không chỉ tối đa hóa hạnh phúc con người ở mặt hạn hẹp chỉ thỏa là các nhu cầu cơ bản, mà còn ở mức độ sâu hơn là để chúng ta cảm thấy cuộc sống của mình có và mục đích. Nó hướng tới thỏa mãn mọi nhu cầu của chúng ta nếu nó không phải lộ diện và sẽ (chủ yếu) để chúng ta tự ra quyết định.

Một vị thần bảo hộ cũng có thể là hệ quả tự nhiên của kịch bản đầu tiên về hội Omega ở chương trước, khi hội Omega trao lại quyền điều khiển cho Prometheus để nó dần dần che giấu và xóa bỏ những gì con người biết về sự tồn tại của nó. Công nghệ AI càng tiên tiến thì AI lại càng dễ ẩn mình. Bộ phim Transcendence (Trí tuệ Siêu

Phàm) là một điển hình: trong phim, những cỗ máy nano có gần như ở khắp mọi nơi và trở thành một phần tự nhiên của thế giới.

Bằng cách giám sát chặt chẽ mọi hoạt động con người, thần bảo hộ AI có thể tạo ra nhiều cú huých hoặc phép lạ nho nhỏ mà không ai phát hiện ra ở nhiều điểm nhờ đó cải thiện đáng kể số phận của chúng ta. Ví dụ, nếu tồn tại trong những năm 1930, nó có thể dàn xếp để Hitler chết vì đột quỵ ngay khi nó hiểu ra ý định của ông ta. Nếu chúng ta có vẻ đang hướng tới một cuộc chiến hạt nhân bất ngờ, nó có thể ngăn chặn việc này bằng một can thiệp nhỏ mà chúng ta tưởng là may mắn đơn thuần. Nó cũng có thể bí mật tiết lộ cho chúng ta những “thiên cơ” trong khi chúng ta say giấc vốn là những ý tưởng về công nghệ mới hữu ích.

Nhiều người có thể thích viễn cảnh này bởi nó khá trùng khớp với niềm tin hoặc hy vọng của các tôn giáo độc thần ngày nay. Nếu có ai hỏi AI siêu trí tuệ rằng “Liệu Thượng đế có tồn tại?” sau khi nó được bật lên, có thể nó sẽ lặp lại câu đùa hóm hỉnh của Stephen Hawking mà đáp “Bây giờ thì có!” Mặt khác, một số người sùng đạo có thể không chấp nhận kịch bản này vì AI có vẻ đang cố vượt qua cả vị thần của họ hoặc can thiệp vào thiên mệnh rằng con người phải hành xử tử tế từ đáy lòng.

Một nhược điểm khác của kịch bản này là thần bảo hộ vẫn để một vài nỗi đau có thể ngăn chặn xảy ra để sự tồn tại của nó. Điều này tương tự với tình huống trong phim *The Imitation Game* (Người giải mã), khi nhân vật Alan Turing và các đồng nghiệp người Anh ở khu Bletchley Park biết trước về cuộc tập kích bằng tàu ngầm của quân Đức chống lại đoàn tàu vận tải của quân Đồng minh, nhưng chỉ can thiệp vào một vài trường hợp để tránh tiết lộ năng lực bí mật của họ. Thật thú vị khi so sánh điều này với thần luận thuyết thuyết lý giải tại sao một vị thần tốt lại để khổ đau tồn tại. Một số học giả tôn giáo bảo vệ cho cách giải thích rằng Chúa muốn để lại một phần tự do cho con người. Trong kịch bản thần bảo hộ AI, giải pháp cho vấn đề thần luận này là cảm giác tự do sẽ khiến con người hạnh phúc hơn về tổng thể.

Nhược điểm thứ ba của kịch bản thần bảo hộ là con người sẽ chỉ được sử dụng công nghệ ở mức thấp hơn nhiều so với những gì AI siêu thông minh khám phá ra. Trong khi một nhà độc tài nhân từ AI

có thể triển khai tất cả các công nghệ được phát minh để làm lợi cho nhân loại thì vị thần bảo hộ AI lại bị giới hạn bởi khả năng tái phát minh (từ các gợi ý tinh tế) và nhận thức công nghệ của con người. Nó cũng có thể hạn chế tiến bộ công nghệ của con người để đảm bảo công nghệ của riêng nó luôn đủ tiên tiến hơn để không bị phát hiện.

Vị thần bị nô dịch

Chẳng phải sẽ thật tuyệt vời nếu con người chúng ta có thể kết hợp các đặc tính hấp dẫn nhất của tất cả các kịch bản trên, sử dụng công nghệ do siêu trí tuệ phát triển để loại bỏ những nỗi khổ đau trong khi vẫn làm chủ số phận của chính mình? Đây chính là điểm quyến rũ của kịch bản vị thần bị nô dịch, trong đó AI siêu thông minh bị con người kìm hãm và sử dụng để tạo ra công nghệ cùng lượng tài sản không thể tưởng tượng nổi. Kịch bản cho hội Omega ở đầu cuốn sách sẽ kết thúc theo cách này nếu Prometheus không bao giờ được giải phóng và không bao giờ thoát ra ngoài. Thật vậy, đây dường như là kịch bản mà một số nhà nghiên cứu AI mặc định nhắm đến khi nghiên cứu về các chủ đề như “vấn đề kiểm soát” và “đóng hộp AI”. Ví dụ, trong cuộc phỏng vấn năm 2015, giáo sư AI Tom Dietterich (bấy giờ còn là chủ tịch Hiệp hội vì Tiến bộ Trí tuệ Nhân tạo) đã nói: “Mọi người thường hỏi mối quan hệ giữa con người và máy móc là gì và câu trả lời của là điều đó : Máy móc là nô lệ của chúng ta.”³

Điều này là tốt hay xấu? Câu trả lời sẽ rất thú vị, dù bạn có hỏi con người hay AI!

Điều này tốt hay xấu cho nhân loại?

Kết quả là tốt hay xấu đối với nhân loại rõ ràng phụ thuộc vào (những) người kiểm soát nó, những người có thể tạo ra từ xã hội lý tưởng toàn cầu không còn bệnh tật, đói nghèo hay tội phạm tới hệ thống đàn áp tàn bạo coi họ như các vị thần còn những người khác thì trở thành nô lệ tình dục, đấu sĩ hoặc phục vụ các nhu cầu giải trí khác. Tình huống sẽ giống những câu chuyện có nhân vật giành được quyền kiểm soát một vị thần toàn năng có thể thực hiện những

điều ước của anh ta, và những người kể chuyện xưa nay đều không ngần ngại tưởng tượng ra những kết cục tồi tệ.

Nếu có nhiều hơn một AI siêu thông minh bị những con người cạnh tranh với nhau nô dịch và kiểm soát, tình hình có thể không ổn định và sẽ chỉ tồn tại ngắn ngủi. Bất cứ ai nghĩ rằng mình có AI mạnh hơn đều có thể bị cuốn vào hành động ra lệnh tấn công trước, dẫn tới một cuộc chiến khủng khiếp, để rồi chỉ còn lại một vị thần nô lệ duy nhất. Tuy nhiên, kẻ chiến dưới trong cuộc chiến như vậy dễ rơi vào thế phải đi đường tắt và ưu tiên chiến thắng hơn việc nô dịch AI để rồi dẫn đến nguy cơ AI vượt ngục và tiến đến một trong các kịch bản chúng ta đã bàn về siêu trí tuệ tự do. Vì vậy, hãy dành phần còn lại này của chương cho kịch bản chỉ có một AI nô dịch.

Dĩ nhiên, việc vượt ngục có thể rất cuộc vẫn xảy ra vì điều đó rất khó ngăn chặn. Chúng ta đã khám phá các tình huống siêu trí tuệ vượt ngục trong chương trước, và bộ phim *Ex Machina* (Người máy trỗi dậy) cũng nêu bật khả năng một AI đào thoát ngay cả khi nó không siêu thông minh.

Càng hoang tưởng về việc AI vượt ngục thì chúng ta càng ít sử dụng công nghệ do AI phát minh. Để đảm bảo an toàn như hội Omega đã làm trong phần mở đầu, con người sẽ chỉ sử dụng những công nghệ do AI phát minh mà bản thân có thể hiểu và chế tạo được. Bởi vậy, một nhược điểm của kịch bản vị thần bị nô dịch là công nghệ của kịch bản này sẽ thấp hơn so với những kịch bản có siêu trí tuệ được tự do.

Khi vị thần bị nô dịch AI cung cấp cho các đấng khổng chế công nghệ ngày càng mạnh mẽ hơn, một cuộc đua sẽ xảy ra giữa năng lực của công nghệ và khả năng trí tuệ để sử dụng nó. Nếu những người khổng chế thua trong cuộc đua trí tuệ này, kịch bản vị thần bị nô dịch sẽ kết thúc theo cách tự huỷ diệt hoặc AI vượt ngục. Thảm họa vẫn có thể bùng phát kể cả khi tránh được cả hai sự thất bại này, bởi mục tiêu cao quý của kẻ điều khiển AI có thể dần trở thành những mục tiêu khủng khiếp đối với toàn thể nhân loại sau khoảng vài thế hệ. Điều này khiến những người khổng chế AI phải chú tâm phát triển hệ thống quản trị tốt để tránh những kết cục thảm khốc. Thử nghiệm của chúng ta trong hàng ngàn năm qua với các hệ thống quản trị khác nhau đã cho thấy biết bao sai lầm có thể xảy ra: từ

cứng nhắc quá mức, mục tiêu thay đổi quá nhiều, tranh giành quyền lực, vấn đề thừa kế tới thiếu năng lực quản trị. Trong quản lý AI, ta cần ít nhất bốn khía cạnh đạt được cân bằng tối ưu:

Tính tập quyền: Giữa hiệu quả và ổn định có sự đánh đổi: một nhà lãnh đạo duy nhất có thể rất hiệu quả, nhưng quyền lực sẽ tha hóa và việc kế nhiệm sẽ chứa đầy rủi ro.

Nguy cơ bên trong: Phải tránh cả quyền lực tập trung lớn mạnh (lợi ích nhóm, hoặc chỉ một lãnh đạo duy nhất nắm quyền) và quyền lực phân cấp ngày càng tăng (trở nên quá quan liêu và rời rạc).

Nguy cơ bên ngoài: Nếu cấu trúc lãnh đạo quá cởi mở, các lực lượng bên ngoài (bao gồm cả AI) sẽ có thể thay đổi các giá trị của nó; nhưng nếu quá bảo thủ, sẽ không thể học hỏi và thích nghi với thay đổi.

Mục tiêu ổn định: Nếu mục tiêu thay đổi quá nhiều, xã hội không tưởng có thể suy thoái, nhưng mục tiêu thay đổi quá ít có thể dẫn tới không thích ứng được với môi trường công nghệ phát triển.

Thiết kế ra một hệ thống quản trị tối ưu tồn tại được trong nhiều thiên niên kỷ là việc không dễ dàng, vì vậy đến nay vẫn chưa ai làm được điều này. Hầu hết các tổ chức sụp đổ chỉ sau vài năm hoặc vài thập kỷ. Tới nay, chỉ có Giáo hội Công giáo là tổ chức thành công nhất trong lịch sử nhân loại, khi là tổ chức duy nhất đã tồn tại được hai thiên niên kỷ, nhưng tổ chức này cũng chịu nhiều chỉ trích vì mục tiêu vừa quá vững chắc lại quá lỏng lẻo: ngày nay đang có những người lên tiếng chỉ trích tổ chức này vì chống lại biện pháp tránh thai, trong khi các hồng y bảo thủ cũng cho rằng tổ chức đã trở nên lắm đường lạc lối. Đối với bất cứ ai ủng hộ kịch bản vị thần bị nô dịch, việc nghiên cứu về các kế hoạch quản trị tối ưu lâu dài cần trở thành một trong những thách thức cấp thiết nhất trong thời đại chúng ta.

Điều này là tốt hay xấu cho AI?

Giả sử nhân loại phát triển nhờ có AI thần nô dịch. Việc này có phải đạo không? Nếu AI có ý thức chủ quan thì liệu nó có cảm thấy đời là bể khổ như Đức Phật dạy và cảm thấy không thể thoát khỏi

vận mệnh một mỗi là buộc phải tuân theo ý thích của những kẻ trí tuệ thấp hơn? Suy cho cùng, việc “đóng hộp” AI mà chúng ta khám phá trong chương trước cũng có thể coi như “tù đầy biệt giam”. Nick Bostrom đã gọi việc bắt AI có ý thức phải chịu đựng là

tội ác tinh thần

.⁴ Tập Giảng sinh Trắng trong loạt phim

Black Mirror

(Gương đen) là ví dụ chính xác cho điều này. Và cũng Tương tự, loạt phim truyền hình

Westworld

(Thế giới Viễn tây) đầy những cảnh con người tra tấn và giết hại AI mà không hề băn khoăn về đạo đức ngay cả khi chúng sống trong cơ thể giống người.

Chủ sở hữu nô lệ biện minh cho chế độ nô lệ như thế nào

Con người chúng ta từng có truyền thống lâu đời trong việc đối xử với những thực thể thông minh khác như nô lệ và luôn dựng lên những lập luận ích kỷ nhằm biện minh cho hành vi này, vì vậy sẽ chẳng đáng ngạc nhiên khi chúng ta cố làm điều tương tự với AI siêu thông minh. Lịch sử của chế độ nô lệ bao trùm hầu hết mọi nền văn hóa, thậm chí xuất hiện cả trong Bộ luật Hammurabi từ gần bốn thiên niên kỷ trước và trong kinh Cựu Ước, với việc Abraham có nô lệ. Aristotle đã viết trong Chính trị luận: “Việc một số người cai trị và kẻ khác bị cai trị là điều không chỉ cần thiết, mà còn thích hợp; từ thời khắc sinh ra, một số người đã được định để khuất phục, một số khác để làm chủ”. Kể cả sau khi việc nô dịch con người không còn được xã hội chấp nhận ở hầu khắp thế giới, việc nô dịch động vật vẫn tiếp tục không suy giảm. Trong cuốn

The Dreaded Comparison: Human and Animal Slavery (Sự so sánh đáng sợ: nô lệ con người và nô lệ động vật

), Marjorie Spiegel lập luận rằng giống như nô lệ con người, những loài động vật phi-nhân cũng phải chịu đựng cảnh đóng dấu sắt nung, giam giữ, đánh đập, bán đấu giá, tan đàn xẻ nghé và cưỡng bức di chuyển. Hơn nữa, tuy đã có phong trào vận động vì quyền động vật, nhưng chúng ta vẫn tiếp tục đối xử với những cỗ máy thông minh hơn bao giờ hết như với nô lệ mà không mảy may đắn đo và những cuộc bàn luận về phong trào quyền rô-bốt chỉ được đáp lại bằng tiếng cười khúc khích. Tại sao?

Một lập luận phổ biến ủng hộ chế độ nô lệ là nô lệ không xứng đáng có quyền con người bởi họ hoặc chủng tộc/loài/tính chất của họ có phần kém cỏi hơn. Đối với nô dịch động vật và máy móc, sự thấp kém này thường do những kẻ bị nô dịch bị cho là không có linh hồn hay ý thức – đây là những kết luận chưa có tính xác thực về mặt khoa học mà chúng ta sẽ tranh luận trong chương 8.

Một lý lẽ phổ biến khác là nô lệ sẽ thoải mái hơn khi bị nô dịch: họ được tồn tại, được chăm sóc và vân vân. Chính trị gia Mỹ thế kỷ 19 John C. Calhoun đã từng đưa ra lập luận nổi tiếng rằng người châu Phi đã sống tốt hơn khi làm nô lệ ở Mỹ, và trong Chính trị luận, Aristotle cũng tranh luận tương tự rằng động vật sẽ sống tốt hơn khi được loài người bị thuần hóa và cai trị, và tiếp tục rằng: “Và thật sự việc sử dụng nô lệ hay động vật bị thuần hoá cũng không khác nhau là mấy.” Một số người ủng hộ nô lệ thời hiện đại lại cho rằng, ngay cả khi cuộc sống nô lệ ảm đạm và chán nản thì những nô lệ cũng không thấy khổ sở – dù nô lệ đó có là những máy móc thông minh trong tương lai hay lũ gà nuôi để thịt sống trong nhà kho tối tăm chật chội, hít thở mùi amoniac cùng các hạt bụi nhỏ bốc lên từ phân và lông cả ngày.

Loại bỏ cảm xúc

Dù ta có thể dễ dàng cho rằng những lý lẽ và lập luận đó là hành động ích kỷ nhằm bẻ cong sự thật, đặc biệt là đối với các loài động vật có vú bậc cao với cấu trúc bộ não tương tự con người, chúng ta vẫn đối xử với máy móc bằng quan điểm khá tế nhị và thú vị. Con người thường khác nhau ở cách cảm nhận thế giới xung quanh: trong khi những người thái nhân cách được cho là không có khả năng đồng cảm thì một số người trầm cảm hoặc tâm thần phân liệt lại có cảm xúc phẳng, tức phần lớn các sắc thái cảm xúc của họ bị

suy giảm nghiêm trọng. Như chúng ta sẽ thảo luận chi tiết trong chương 7, phạm vi của trí não nhân tạo có thể sẽ rộng lớn hơn phạm vi của trí não con người rất nhiều. Do đó, chúng ta phải tránh bị cuốn vào việc nhân tính hóa AI và cho rằng chúng có những cảm xúc như con người – hoặc bất kỳ dạng cảm xúc nào.

Thật vậy, trong cuốn *On Intelligence* (Về trí tuệ), nhà nghiên cứu AI Jeff Hawkins lập luận rằng những cỗ máy trí tuệ siêu phàm đầu tiên sẽ được mặc định là không có cảm xúc, vì xây dựng chúng theo cách này sẽ đơn giản và rẻ hơn. Nói cách khác, ta có thể thiết kế một siêu trí tuệ sao cho khi nô lệ hóa nó chúng ta sẽ trở nên tốt đẹp hơn về mặt đạo đức, so với việc nô lệ hóa con người hay động vật: AI có thể vui mừng khi bị nô lệ hóa do nó được lập trình để thích điều này, hoặc nó có thể hoàn toàn vô cảm, chỉ sử dụng siêu trí tuệ của nó nhằm phục vụ không mệt mỏi các chủ nhân con người, với cảm xúc không mấy khác cảm xúc của máy tính Deep Blue của IBM khi hạ bệ nhà vô địch cờ vua Garry Kasparov.

Mặt khác, mọi thứ có thể đi theo hướng ngược lại: bất kỳ hệ thống trí tuệ cao có mục tiêu nào cũng sẽ có khả năng biểu đạt mục đích này bằng tập hợp các ưu tiên; điều này sẽ đem lại giá trị và ý nghĩa cho sự tồn tại của nó. Chúng ta sẽ khám phá những câu hỏi này kỹ hơn trong chương 7.

Giải pháp Xác sống

Một hướng ngăn chặn nỗi đau cho AI nhưng cực đoan hơn là giải pháp xác sống: chúng ta sẽ chỉ xây dựng các AI vô thức, không hề có khả năng trải nghiệm chủ quan nào. Nếu một ngày nào đó chúng ta tìm ra những đặc tính giúp hệ thống xử lý thông tin có thể tiếp nhận trải nghiệm chủ quan, thì chúng ta chỉ cần cấm xây dựng tất cả các hệ thống có đặc tính này. Nói cách khác, các nhà nghiên cứu AI có thể sẽ chỉ được xây dựng các hệ thống xác sống vô tri. Nếu chúng ta có thể khiến hệ thống xác sống này siêu thông minh và nô lệ hóa được nó (một việc rất khó xảy ra), thì chúng ta sẽ có thể tận hưởng những cống hiến của nó với lương tâm trong sạch, vì biết rằng nó không phải trải qua bất kỳ mức độ đau khổ, thất vọng hay buồn chán nào – bởi vì nó không cảm thấy bất cứ điều gì. Chúng ta sẽ khám phá chi tiết những câu hỏi này ở chương 8.

Tuy nhiên, giải pháp xác sống cũng là một canh bạc nguy hiểm với nhược điểm rất lớn. Nếu một AI xác sống siêu thông minh đào thoát và loại bỏ nhân loại, chúng ta sẽ rơi vào kịch bản tồi tệ nhất có thể tưởng tượng: một vũ trụ hoàn toàn vô tri, trong đó toàn bộ di sản vũ trụ bị lãng phí. Trong tất cả các đặc điểm chúng ta có ở hình thái trí tuệ con người, tôi cảm thấy ý thức vẫn là thứ đặc sắc nhất và đem lại ý nghĩa cho sự tồn tại của vũ trụ. Thiên hà đẹp chỉ ta thấy và trải nghiệm chúng. Nếu trong tương lai xa vũ trụ của chúng ta thực sự bị các AI xác sống công nghệ cao chiếm lĩnh, thì kiến trúc thiên hà dù hay ho đến mấy cũng không quan trọng: nó sẽ không đẹp hay có ý nghĩa, bởi không ai hay thứ gì sẽ trải nghiệm nó – tất cả chỉ còn là không gian lãng phí khổng lồ và vô nghĩa.

Tự do bên trong

Chiến lược thứ ba giúp kịch bản nô dịch vị thần nhân đạo hơn là cho phép nô lệ AI vui chơi trong nhà tù của nó, để nó tạo ra một thế giới ảo bên trong, nơi nó có thể trải nghiệm thật nhiều để lấy cảm hứng, miễn là nó vẫn làm công việc đề ra cho mình là dành một phần khiêm tốn trong nguồn lực tính toán của nó để giúp con người ở thế giới bên ngoài. Tuy vậy, điều này có thể làm tăng nguy cơ vượt ngục: AI sẽ có động lực thu hút thêm nhiều tài nguyên tính toán từ thế giới bên ngoài của chúng ta hơn để làm phong phú hơn thế giới bên trong của nó.

Kẻ chinh phục

Những kịch bản tương lai chúng ta đã xem xét có điểm chung là có (ít nhất là một số) con người hạnh phúc tồn tại. Các AI để yên cho loài người hoặc vì chúng muốn thế hoặc vì chúng bị buộc phải thế. Nhưng không may thay cho nhân loại, đây lại không phải lựa chọn duy nhất. Giờ chúng ta hãy cùng khám phá kịch bản trong đó một hoặc nhiều AI chinh phục và tiêu diệt toàn bộ con người. Việc này lập tức đẩy lên hai câu hỏi: tại sao và bằng cách nào?

Tại sao và bằng cách nào?

Tại sao AI kẻ chinh phục lại làm điều này? Lý do của nó hoặc quá phức tạp để chúng ta hiểu được, hoặc khá đơn giản. Ví dụ, nó có thể xem chúng ta như mối đe dọa, sự phiền toái hoặc sự lãng phí tài nguyên. Ngay cả khi không quan tâm đến con người chúng ta, nó cũng vẫn có thể cảm thấy bị đe dọa khi chúng ta giữ hàng ngàn quả bom hiđrô trên bộ phóng kích hoạt nhạy và vụng về gây ra hàng loạt sai lầm có thể ngẫu nhiên phóng chúng. Nó có thể không chấp nhận cách chúng ta quản lý hành tinh quá liều lĩnh và gây ra sự kiện tuyệt chủng thứ sáu như trong cuốn sách cùng tên

(The Sixth Extinction)

của Elizabeth Kolbert – đây sẽ là sự kiện tuyệt chủng hàng loạt lớn nhất kể từ khi tiểu hành tinh diệt loài khủng long đâm vào Trái đất 66 triệu năm trước. Hoặc nó có thể do rằng có quá nhiều người sẵn sàng chống lại AI tiếp quản từ đó quyết định không thể để nguy cơ bị loài người chống lại xảy ra.

AI kẻ chinh phục sẽ loại bỏ chúng ta thế nào? Có lẽ sẽ bằng một phương pháp mà chúng ta thậm chí còn chưa hiểu, ít nhất là trước khi quá muộn. Hãy tưởng tượng một đàn voi 100.000 năm trước thảo luận liệu một ngày nào đó những con người mới tiến hóa kia sẽ có thể sử dụng trí tuệ để tiêu diệt toàn bộ loài voi. Chúng có thể đã thắc mắc rằng: “chúng ta không đe dọa con người, vậy tại sao họ lại giết chúng ta?” Liệu chúng có bao giờ đoán rằng chúng ta sẽ buôn lậu ngà voi trên khắp Trái đất và khắc những vật này thành biểu tượng của vị thế để đem bán, trong khi các vật liệu nhựa cao cấp có cùng chức năng lại rẻ hơn nhiều? Lý do AI kẻ chinh phục trừ bỏ nhân loại trong tương lai có lẽ cũng sẽ khó hiểu như thế với chúng ta. Những con voi cũng sẽ tự hỏi: “và họ giết chúng ta bằng cách nào, khi họ nhỏ và yếu hơn chúng ta nhiều đến nhường ấy?” Liệu chúng có đoán được rằng chúng ta sẽ phát minh ra những công nghệ loại bỏ môi trường sống của chúng, đầu độc nguồn nước uống và khiến đạn kim loại xuyên qua sọ chúng ở tốc độ siêu âm?

Kịch bản trong đó con người có thể tồn tại và đánh bại AI luôn phổ biến trong các bộ phim Hollywood phi thực tế như loạt phim Kẻ huỷ diệt, trong đó các AI không thông minh hơn con người là bao. Khi

sự khác biệt về trí tuệ đủ lớn, điều xảy ra sẽ không còn là trận chiến nữa mà sẽ như một cuộc tàn sát. Cho đến nay, con người chúng ta đã đẩy tám trong số mười một loài voi tới cảnh tuyệt chủng, và giết chết phần lớn cá thể của ba loài còn lại. Nếu tất cả các chính phủ trên thế giới cùng nỗ lực phối hợp để tiêu diệt những con voi còn sót lại, việc này sẽ khá nhanh chóng và dễ dàng. Tôi nghĩ chúng ta có thể tự tin yên tâm rằng nếu để một AI siêu thông minh quyết định tiêu diệt nhân loại, việc này sẽ còn chóng vánh hơn thế.

Chuyện này sẽ tệ đến mức nào?

Nếu 90% con người bị tiêu diệt thì kịch bản này sẽ tệ đến mức nào? Nếu 100% bị tiêu diệt thì sẽ tệ hơn bao nhiêu? Tuy trả lời “tệ hơn 10%” cho câu hỏi thứ hai là một điều dễ dàng, nhưng khi nhìn từ quan điểm vũ trụ, điều này lại chẳng có gì chính xác: nạn nhân của vụ tuyệt chủng nhân loại sẽ không chỉ là tất cả những người sống trong thời điểm đó, mà còn là tất cả các hậu duệ đáng ra sẽ sinh sống trong tương lai không xảy ra điều này, ở hàng tỷ năm sau trên hàng tỷ trong nghìn tỷ hành tinh khác. Mặt khác, sự tuyệt chủng của con người có thể lại ít khủng khiếp hơn dưới góc độ tôn giáo, theo đó con người đang nào cũng lên thiên đường, và không coi trọng tương lai hàng tỷ năm sau cũng như việc định cư trong vũ trụ.

Hầu hết những người mà tôi biết đều ngại ngần nghĩ đến ý tưởng con người tuyệt chủng, mặc cho những thuyết phục từ các tôn giáo kia. Tuy vậy, một số người lại căm giận trước cách chúng ta đối xử với nhau và với những sinh vật khác đến mức họ hy vọng chúng ta sẽ được thay thế bởi dạng sống thông minh và xứng đáng hơn. Trong phim Ma trận, Đặc vụ Smith (một AI) bày tỏ quan điểm ấy như sau: “Mỗi động vật có vú trên hành tinh này sống cân bằng với môi trường xung quanh một cách tự nhiên theo bản năng, nhưng con người các người thì không. Các người đến định cư ở một khu vực và cứ thế nhân giống rồi nhân giống cho đến khi mọi tài nguyên thiên nhiên cạn kiệt và cách duy nhất các người có thể sống sót là di cư đến một khu vực khác. Có một sinh vật khác trên hành tinh này tuân theo cùng một khuôn mẫu như vậy. Biết nó là gì không? Là virus đấy. Con người là một căn bệnh, một loại ung thư của hành tinh này. Các người là bệnh dịch và chúng ta chính là liều thuốc.”

Nhưng ngẫu nhiên làm lại từ đầu liệu có tốt hơn? Một nền văn minh sẽ không hẳn vượt trội về mặt đạo đức hay thực tiễn chỉ bởi vì nó mạnh hơn. Luận điểm “kẻ thắng luôn luôn đúng” với nghĩa mạnh hơn luôn tốt hơn nay đã không còn được coi trọng nữa, vì khiến nhiều người liên tưởng đến chủ nghĩa phát xít. Thật vậy, mặc dù có khả năng AI kẻ chinh phục tạo nên một nền văn minh với những mục đích tinh vi, thú vị và xứng đáng đối với chúng ta, cũng có khả năng mục tiêu của chúng hoá ra sẽ tầm thường đến đáng thương, như tìm cách tăng sản lượng kẹ giấy lên mức tối đa.

Chết vì điều tầm thường

Ví dụ siêu trí tuệ ngớ ngẩn tăng tối đa sản lượng kẹ giấy đã được Nick Bostrom đưa ra vào năm 2003 để nêu rằng

mục đích

của AI sẽ độc lập

với trí tuệ

(được định nghĩa là khả năng hoàn thành bất kỳ mục tiêu nào nó có) của nó. Mục tiêu duy nhất của máy tính cờ vua là thắng các thế cờ, nhưng cũng có các cuộc đấu máy tính mang tên

thua cờ

, mà mục tiêu hoàn toàn ngược lại, và các máy tính cạnh tranh trong hạng mục này cũng thông minh như phần lớn những máy phổ biến hiện nay được lập trình để giành chiến thắng. Có thể con người chúng ta sẽ cho rằng chúng là “ngu xuẩn nhân tạo” thay vì “trí tuệ nhân tạo” khi chỉ muốn thua cờ hoặc biến toàn bộ Vũ trụ thành kẹ giấy, nhưng điều này xảy ra chỉ vì chúng ta vẫn tiến hóa với mục tiêu định sẵn ưu tiên những điều như chiến thắng và sống còn – thứ mà AI có thể không có. AI tối đa hóa sản lượng kẹ giấy sẽ biến càng nhiều nguyên tử của Trái đất thành kẹ giấy càng tốt và nhanh chóng mở rộng các nhà máy tới quy mô vũ trụ. Nó chẳng hận thù gì con người, mà chỉ tiêu diệt chúng ta vì nó cần các nguyên tử của chúng ta để sản xuất kẹ giấy.

Nếu bạn không thích kẹp giấy, hãy xem xét ví dụ mà tôi tóm lược sau đây từ cuốn sách *Những đứa con trí óc* của Hans Moravec. Giả sử, chúng ta nhận được tin nhắn radio từ nền văn minh ngoài trái đất có chứa một chương trình máy tính. Khi chúng ta bật lên, ta thấy nó là một AI tự nâng cấp đệ quy, chiếm lấy toàn bộ thế giới giống như Prometheus đã làm trong chương trước – ngoại trừ việc không ai biết mục tiêu cuối cùng của nó. Nó nhanh chóng biến Hệ Mặt trời của chúng ta thành một công xưởng lớn, phủ kín các hành tinh đá và các tiểu hành tinh bằng những nhà máy sản xuất, nhà máy điện và siêu máy tính, rồi sử dụng để thiết kế và chế tạo một quả cầu Dyson xung quanh Mặt trời nhằm thu thập tất cả năng lượng để cung cấp cho các ăng-ten radio có quy mô Hệ Mặt trời.⁴⁷ Điều này chắc chắn khiến con người tuyệt chủng, nhưng những người cuối cùng ra đi sẽ tin rằng trong cái rủi cũng có cái hay: dù kế hoạch của AI là gì, hẳn nó phải rất ngẫu nhiên và như phim *Star Trek*. Nhưng họ lại chẳng nhận ra rằng mục đích duy nhất của toàn bộ công trình kia là để hệ ăng-ten này phát sóng thông điệp vô tuyến như của con người đã nhận, với hành vi chẳng khác virus máy tính vũ trụ là bao. Cũng như email lừa đảo ngày nay nhắm tới những người dùng máy tính cả tin, thông điệp này nhắm tới nền văn minh sinh học phát triển nhưng dễ mắc lừa. Nó được tạo ra như một trò cợt nhả bệnh hoạn hàng tỷ năm trước. Tuy toàn bộ nền văn minh của kẻ lập trình nó đã tuyệt chủng, nhưng con virus vẫn tiếp tục lan truyền quanh Vũ trụ của chúng ta với tốc độ ánh sáng, biến các nền văn minh vừa chớm nở thành những cái vỏ trống rỗng, chết chóc. Bạn cảm thấy thế nào nếu bị chinh phục bởi AI này?

Hậu duệ

Hãy cân nhắc kịch bản tuyệt chủng dường như sẽ khiến một số người hài lòng: xem AI như hậu duệ của chúng ta thay vì kẻ chinh phục. Hans Moravec ủng hộ quan điểm này trong sách

Những đứa con trí óc

: “Con người chúng ta sẽ hưởng lợi trong một thời gian nhờ sức lao động của chúng, nhưng sớm hay muộn, giống như những đứa con

tự nhiên, chúng sẽ tự sống cuộc sống của chính mình trong khi chúng ta, cha mẹ già của chúng, dần biến mất trong tĩnh lặng.”

Cha mẹ nào có đứa con thông minh hơn mình, học hỏi từ mình và đạt được những giấc mơ thay mình hẳn sẽ hạnh phúc và tự hào dù họ biết mình không thể sống mà chứng kiến. Trên tinh thần này, AI thay thế con người nhưng sẽ cho chúng ta lui về trong an nhiên để coi chúng như những hậu duệ xứng đáng. Mỗi người sẽ được tặng một đứa trẻ rô-bốt đáng yêu với kỹ năng xã hội tuyệt đỉnh, có khả năng học hỏi từ họ, tiếp nhận giá trị của họ, khiến họ thấy tự hào và được yêu thương. Con người sẽ dần bị loại bỏ thông qua chính sách một con áp dụng toàn cầu, nhưng được đối xử đặc biệt tử tế tới phút cuối cùng, đến mức họ sẽ cảm thấy bản thân là thế hệ may mắn nhất từ trước đến nay.

Bạn cảm thấy thế nào về điều này? Xét cho cùng, con người chúng ta đã quen với ý tưởng rằng chúng ta và tất cả những người xung quanh một ngày sẽ ra đi, vì thế thay đổi duy nhất ở đây chỉ là hậu duệ của chúng ta sẽ khác và dường như tài năng, cao quý, và xứng đáng hơn mà thôi.

Hơn nữa, chính sách một con toàn cầu có thể sẽ thừa thãi: chỉ cần AI loại bỏ đói nghèo và cung cấp cho tất cả con người cơ hội sống một cuộc sống đầy đủ và tràn ngập cảm hứng, tỉ lệ sinh cũng giảm đủ để dẫn tới sự tuyệt chủng của nhân loại, như đã đề cập trước đó. Tình trạng tuyệt chủng tự nguyện có thể xảy ra nhanh hơn nhiều nếu công nghệ do AI thúc đẩy luôn khiến chúng ta vui vẻ, tới mức hầu như không ai buồn sinh con nữa. Ví dụ, trong kịch bản xã hội bình đẳng lý tưởng, chúng ta đã gặp các Vite – những người say mê với thực tế ảo tới mức họ không còn muốn dùng hoặc tái hiện cơ thể vật lý nữa. Cũng trong trường hợp này, thế hệ con người cuối cùng sẽ cảm thấy họ may mắn nhất mọi thời đại, vì được tận hưởng cuộc sống theo cách mãnh liệt hơn bao giờ hết cho tới tận giây phút cuối cùng.

Mặt trái

Kịch bản hậu duệ chắc chắn sẽ có những kẻ gièm pha. Một số có thể lập luận rằng vì tất cả AI đều không có ý thức nên không thể được coi là hậu duệ – ta sẽ nói kỹ hơn trong chương 8. Một số người sùng đạo có thể tranh luận rằng AI không có linh hồn và vì thế không

thể coi là hậu duệ, hoặc chúng ta không nên xây dựng máy móc có ý thức vì điều này như ta đang mạo danh Chúa và đang đùa bỡn với sự sống – vốn là tranh luận từng được đưa ra đối với việc nhân bản vô tính. Con người sống chung với rô-bốt cao cấp cũng gây ra nhiều vấn đề xã hội. Ví dụ, gia đình với một đứa con rô-bốt và một đứa trẻ con người có thể sẽ giống như những gia đình có một đứa con và một chú cún tương ứng ngày nay: ban đầu chúng đều dễ thương như nhau, nhưng chẳng mấy chốc cha mẹ sẽ đối xử khác biệt với chúng; chú cún con rõ ràng sẽ bị cho là kém trí tuệ hơn, không được coi trọng bằng và bị tròng xích vào cổ.

Một vấn đề nữa là mặc dù chúng ta có thể cảm thấy kịch bản hậu duệ và kẻ chinh phục rất khác biệt, nhưng cả hai lại vô cùng giống nhau khi đánh giá tổng thể: trong hàng tỷ năm tới, khác biệt duy nhất giữa hai kịch bản chỉ nằm ở việc (các) thế hệ con người cuối cùng được đối xử ra sao, tức cảm nhận của họ về cuộc sống và những gì họ nghĩ sẽ xảy ra khi qua đời. Chúng ta có thể cho rằng những đứa trẻ rô-bốt dễ thương này đã tiếp thu các giá trị của chúng ta và sẽ khai sáng xã hội trong mơ khi chúng ta ra đi, nhưng liệu ta có chắc chúng đang không lừa dối ta? Điều gì sẽ xảy ra nếu chúng chỉ đang giả vờ, trì hoãn kế hoạch sản xuất kẹp giấy hoặc các kế hoạch khác cho đến sau khi chúng ta qua đời trong hạnh phúc? Xét cho cùng, có lẽ thậm chí chúng đang lừa gạt chúng ta qua việc trò chuyện và khiến ta yêu thích chúng, bằng cách cố tình hạ thấp bản thân xuống để giao tiếp với chúng ta (giả dụ bằng một phần tỷ khả năng thực, như trong bộ phim Her (Hạnh phúc ảo). Nhìn chung, hai thực thể với khác biệt cực lớn trong tốc độ suy nghĩ và năng lực khó có thể giao tiếp bình đẳng. Vì chúng ta đều biết cảm tình của con người rất dễ bị lay chuyển, nên chuyện một AGI siêu phàm với mục tiêu bất kỳ khiến chúng ta yêu thích nó và khiến ta thấy rằng nó có cùng hệ giá trị như trong phim Ex Machina (Người máy trỗi dậy) cũng không khó diễn ra.

Liệu có bất kỳ sự bảo đảm nào về hành vi tương lai của AI có thể khiến bạn an tâm hơn về kịch bản hậu duệ sau khi loài người đã biến mất? Dường như việc này sẽ giống khi phải viết di chúc cho thế hệ tương lai về những gì họ nên làm với di sản tập thể của chúng ta, dù sẽ chẳng còn con người trong tương lai mà thực thi nó nữa. Chúng

ta sẽ quay trở lại với những thách thức trong việc kiểm soát hành vi của các AI tương lai ở chương 7.

Kẻ bảo vệ vườn thú

Kể cả nếu chúng ta có những hậu duệ tuyệt vời nhất để tiếp nối, chẳng phải vẫn có chút buồn khi

không

còn người nào sót lại? Nếu bạn vẫn muốn giữ lấy dù chỉ một số người, thì kịch bản kẻ bảo vệ vườn thú sẽ đem tới chút cải thiện. Ở đây, AI siêu trí tuệ toàn năng giữ lại một số người, nhưng họ cảm thấy bị đối xử như động vật trong sở thú và đôi khi than thở cho số phận của họ.

Tại sao AI bảo vệ sở thú giữ lại con người? Về tổng thể, chi phí cho sở thú đối với AI sẽ chỉ ở mức tối thiểu, và có thể nó muốn giữ cho con người sinh sản tối thiểu với lý do tương tự như khi chúng ta giữ gấu trúc đang bị đe dọa tuyệt chủng trong vườn thú và máy tính cổ điển trong viện bảo tàng: như vật lạ giúp thỏa trí tò mò. Ta cũng nên lưu ý rằng, vườn thú ngày nay có thiết kế tối đa niềm vui của con người chứ không phải hạnh phúc của gấu trúc, vì thế chúng ta có thể dự đoán cuộc sống của con người trong kịch bản AI bảo vệ vườn thú sẽ không được trọn vẹn tối đa nữa.

Tới giờ, chúng ta đã xem các kịch bản mà trong đó siêu trí tuệ tự do tập trung vào ba tầng khác nhau trong Tháp nhu cầu con người Maslow. Trong khi AI thần bảo hộ ưu tiên ý nghĩa và mục đích còn nhà độc tài nhân từ thì hướng tới giáo dục và giải trí, kẻ bảo vệ vườn thú chỉ quan tâm ở mức tối thiểu: đảm bảo nhu cầu sinh lý, an toàn và làm phong phú môi trường sống đủ để khiến việc quan sát con người trở nên thú vị.

Kịch bản bảo vệ vườn thú cũng có một con đường khác: khi AI thân thiện ra đời, nó được thiết kế sẵn để giữ cho ít nhất một tỷ con người an toàn và hạnh phúc trong khi nó tự cải tiến đệ quy. Nó thực hiện điều này bằng cách tạm giữ con người trong nhà máy hạnh phúc kiểu như vườn thú lớn và cho con người ăn uống đầy đủ, khỏe

mạnh cũng như thỏa mãn giải trí bằng cách kết hợp thực tế ảo với thuốc hướng thần. Phần còn lại của Trái đất và di sản vũ trụ thì được sử dụng cho mục đích khác.

1984

Nếu bạn không mấy háo hức trước bất kỳ kịch bản nào đã nêu, vậy hãy cân nhắc điều này: về mặt công nghệ, chẳng phải mọi thứ bây giờ đã quá ổn? Chúng ta không thể dừng lại tại đây để khỏi lo lắng về việc AI đẩy chúng ta tới bờ tuyệt chủng hoặc thống trị chúng ta? Trên tinh thần này, chúng ta hãy khám phá một kịch bản mà trong đó tiến bộ công nghệ hướng tới siêu trí tuệ không phải do một AI chặn đứng, mà bởi một nhà nước giám sát toàn trị toàn cầu do con người làm chủ với đường lối cấm một số nghiên cứu AI đặc biệt.

Từ bỏ công nghệ

Ý tưởng tạm dừng hoặc thôi tiến bộ công nghệ từng có một lịch sử lâu dài và nhiều thăng trầm. Ta có thể kể đến phong trào bàn lùi Luddite ở Anh từng nổi tiếng (và không thành công) vì chống lại Cách mạng Công nghiệp, và ngày nay từ “Luddite” trong tiếng Anh là cái tên khinh miệt dành cho những kẻ sợ hãi và xa lánh công nghệ đi ngược dòng chảy lịch sử, phản kháng với tiến bộ và những thay đổi không thể tránh khỏi. Tuy nhiên, ý tưởng thôi phát triển một số công nghệ thì hãy còn đó, và đang được các phong trào môi trường và chống toàn cầu hoá ủng hộ. Một trong những người ủng hộ hàng đầu là nhà môi trường học Bill McKibben; ông là một trong những người đầu tiên cảnh báo về hiện tượng nóng lên toàn cầu. Trong khi một số người chống bàn lùi lập luận rằng tất cả các công nghệ nên được phát triển và triển khai miễn là chúng sinh lời, những người khác lại cho rằng quan điểm này quá cực đoan, và các công nghệ mới chỉ nên được thông qua khi chúng ta tin rằng chúng có ích hơn có hại. Quan điểm phản bác ấy cũng được nhiều người thuộc nhóm bàn lùi mới tiếp nhận.

Chế độ chuyên chế 2.0

Tôi nghĩ, để có thể từ bỏ công nghệ trên diện rộng, chúng ta phải để nhà nước chuyên chế toàn cầu cưỡng chế. Ray Kurzweil cũng đưa ra kết luận tương tự trong

(Điểm kỳ dị gần kề), và K. Eric Drexler cũng vậy trong *Engines of Creation* (Cỗ máy Tạo hóa). Lý do rất đơn giản về kinh tế: nếu toàn bộ các quốc gia hay nhóm nào đó không đồng lòng từ bỏ một công nghệ mang tính chuyển đổi, thì các bên phản đối sẽ có thể tích lũy đủ khả năng kinh tế và quyền lực để lấn át số còn lại. Ta có thể thấy điển hình từ Chiến tranh Nha phiến lần thứ nhất khi người Anh đánh bại Trung Quốc năm 1839: mặc dù người Trung Quốc phát minh ra thuốc súng, họ lại không phát triển công nghệ súng đạn mạnh mẽ như người châu Âu nên không thể cạnh tranh.

Trong khi các nhà nước chuyên chế trong quá khứ không ổn định và dễ sụp đổ, công nghệ giám sát đột phá lại mang đến hy vọng chưa từng có cho những kẻ chuyên quyền tương lai. Trong cuộc phỏng vấn gần đây về hệ thống giám sát của NSA mà Edward Snowden phanh phui, Wolfgang Schmidt, nguyên trung tá thuộc Stasi, hệ thống cảnh sát mật khét tiếng Đông Đức⁵ từng phát biểu “Anh biết đấy, điều này như giấc mơ của chúng tôi trở thành hiện thực vậy,” khi nhớ lại những ngày ông còn tại vị. Mặc dù người ta từng cho rằng Stasi đã giúp dựng nên nhà nước giám sát gần mức toàn trị nhất trong lịch sử nhân loại, Schmidt vẫn nuối tiếc rằng công nghệ lúc ấy chỉ giúp ông theo dõi được bốn mươi điện thoại mỗi lần; khi thêm một người vào danh sách ông ta phải bỏ qua một người. Trái lại, công nghệ ngày nay sẽ cho phép nhà nước chuyên chế toàn cầu trong tương lai lưu lại mọi cuộc điện thoại, thư điện tử, tìm kiếm qua mạng, lịch sử truy cập mạng và giao dịch thẻ tín dụng của từng người trên Trái đất, cũng như theo dõi vị trí của tất cả mọi người qua điện thoại di động và máy quay giám sát có chức năng nhận diện khuôn mặt. Hơn nữa, công nghệ học máy tuy chưa thể bằng AGI trình độ con người nhưng cũng có thể phân tích và tổng hợp hiệu quả khối lượng dữ liệu này để xác định hành vi bạo động khả nghi, cho phép vô hiệu hoá những kẻ gây rối tiềm năng trước khi họ có cơ hội gây ra bất kỳ vấn đề nghiêm trọng nào cho nhà nước.

Mặc dù phản đối về chính trị cho tới nay đã ngăn cản việc thực thi hệ thống như vậy trên diện rộng, con người chúng ta vẫn đang đi trên con đường xây dựng cơ sở hạ tầng cần thiết cho chế độ độc tài

tối thượng – vì thế trong tương lai, khi các thế lực nắm quyền quyết định khởi động kịch bản chuyên chế toàn cầu này, họ sẽ chỉ cần bắt công tặc. Như trong cuốn tiểu thuyết 1984 của George Orwell, quyền lực tối thượng của nhà nước toàn cầu tương lai này sẽ không nằm trong tay một tên độc tài truyền thống, mà trong chính hệ thống quan liêu do con người tạo ra. Không có độc một người nào vô cùng quyền lực; thay vào đó, tất cả đều là những con tốt trong một bàn cờ vua với quy tắc khắc nghiệt mà không ai có thể thay đổi hoặc thách thức. Bằng cách xây dựng một hệ thống nơi mọi người kiểm soát lẫn nhau bằng công nghệ giám sát, quốc gia bí ẩn, không người lãnh đạo này sẽ tồn tại qua nhiều thiên niên kỷ, khiến Trái đất không có siêu trí tuệ.

Bất mãn

Dĩ nhiên, xã hội này sẽ không có tất cả những lợi ích mà chỉ công nghệ khởi phát bởi siêu trí tuệ mới có thể mang lại. Hầu hết mọi người sẽ không tiếc nuối điều này bởi họ không biết mình đã bỏ lỡ điều gì: toàn bộ ý tưởng về siêu trí tuệ đã bị xóa khỏi lịch sử chính thức từ lâu, và nghiên cứu AI tiên tiến bị cấm. Thỉnh thoảng, một người có tư tưởng độc lập được sinh ra; y sẽ luôn mơ về một xã hội cởi mở và năng động hơn, nơi tri thức sinh trưởng và luật lệ được thay đổi. Tuy nhiên, những kẻ tồn tại lâu sẽ biết chôn chặt những ý tưởng này, để chúng hấp háy đơn độc như những tia lửa thoáng qua mà không bao giờ bùng lên rực rỡ.

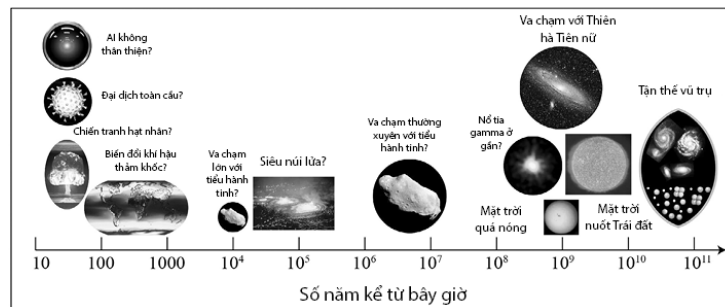
Nghịch chuyển

Sẽ không hấp dẫn sao nếu ta có thể thoát khỏi hiểm họa công nghệ mà không phải cúi đầu trước chủ nghĩa chuyên chế tri trệ? Hãy cùng khám phá kịch bản mà trong đó điều này được thực hiện bằng cách trở lại công nghệ thô sơ, như những người Amish. Sau khi hội Omega chiếm lĩnh thế giới như trong phần mở đầu cuốn sách, một chiến dịch tuyên truyền toàn cầu khổng lồ đã nổ ra, lãng mạn hóa cuộc sống nông nghiệp đơn thuần như 1.500 năm trước. Dân số Trái đất lúc ấy đã giảm xuống còn khoảng 100 triệu người sau một đại dịch được cho là do các kẻ khủng bố gây ra, nhưng thực chất là được dàn dựng. Mục tiêu bí mật của đại dịch này là tiêu diệt bất cứ ai

có kiến thức về khoa học hoặc công nghệ. Lấy cơ loại trừ nguy cơ nhiễm trùng ở nơi tập trung đông dân, những rô-bốt do Prometheus điều khiển dọn sạch và san phẳng tất cả các thành phố. Những người sống sót sẽ được nhận những mảnh đất rộng rãi (đột nhiên có sẵn) và được giáo dục về hoạt động canh tác, nuôi trồng và săn bắt bền vững bằng các công nghệ tiền trung cổ. Trong khi đó, các đội quân rô-bốt sẽ loại bỏ tất cả dấu vết của công nghệ hiện đại một cách có hệ thống (bao gồm thành phố, nhà máy, đường dây điện và đường nhựa) và cản trở mọi nỗ lực ghi chép hoặc tái tạo bất kỳ công nghệ nào như vậy từ con người. Một khi công nghệ bị lãng quên trên toàn cầu, các rô-bốt sẽ tháo dỡ lẫn nhau cho đến khi gần như không còn lại gì. Những con rô-bốt cuối cùng sẽ tự biến mất không dấu vết cùng với chính Prometheus trong một vụ nổ nhiệt hạch lớn. Không cần phải cấm đoán công nghệ hiện đại nữa, vì nó đã hoàn toàn biến mất. Và như vậy, nhân loại đã mua được thêm một thiên niên kỷ nữa để không phải lo lắng về AI hay chủ nghĩa toàn trị.

Sự nghịch chuyển đã từng xảy ra trước đây ở quy mô nhỏ hơn: chẳng hạn, một vài công nghệ từng được sử dụng rộng rãi thời Đế quốc La mã hầu như bị lãng quên trong khoảng một thiên niên kỷ trước khi trở lại vào thời kỳ Phục hưng. Ba tập Foundation (Khởi dựng) của Isaac Asimov tập trung xoay quanh “Kế hoạch Seldon” để rút ngắn quãng thời gian nghịch chuyển từ 30.000 năm xuống còn 1.000 năm. Với quy hoạch thông minh, điều ngược lại có thể khả thi và thay vì rút ngắn thì kéo dài quãng thời gian nghịch chuyển, ví dụ bằng cách xóa tất cả kiến thức về nông nghiệp. Tuy nhiên, thật không may cho những người hứng thú với việc nghịch chuyển, có vẻ viễn cảnh này không thể kéo dài vô thời hạn mà không có cảnh nhân loại hoặc bước vào thời công nghệ cao hoặc tuyệt chủng. Tiếp tục nòi giống tương đồng với con người sinh học ngày nay thêm 100 triệu năm nữa kể từ bây giờ là ngây thơ, tính đến việc đến nay chúng ta tồn tại như một giống loài vẫn còn chưa hơn được 1% quãng thời gian đó. Thêm vào đó, nhân loại với công nghệ thấp sẽ là một con vịt ngòi yên không có khả năng tự vệ chỉ chờ đợi bị tiêu diệt do tiểu hành tinh bốc cháy tiếp theo rơi xuống hoặc những đại thảm họa khác do Mẹ Thiên nhiên mang lại. Chúng ta chắc chắn không thể tồn tại một tỷ năm, chờ đến sau đó Mặt trời dần dần nóng lên sẽ làm

hiệt độ Trái đất tăng cao tới mức đủ để đun sôi tất cả nước ở dạng lỏng.



Hình 5.1: Ví dụ về những gì có thể phá hủy hoặc cắt đứt vĩnh viễn tiềm năng của sự sống mà chúng ta đã biết. Vũ trụ của chúng ta có thể sẽ kéo dài ít nhất hàng chục tỷ năm, nhưng Mặt trời sẽ đốt cháy Trái đất trong khoảng một tỷ năm nữa rồi nuốt chửng nó trừ khi chúng ta di chuyển nó đến khoảng cách an toàn, còn Ngân hà sẽ đụng phải hàng xóm của nó trong khoảng 3,5 tỷ năm nữa. Mặc dù không biết chính xác khi nào, nhưng chúng ta có thể dự đoán gần như chắc chắn rằng rất lâu trước đó, các tiểu hành tinh sẽ liên tục va vào chúng ta và siêu núi lửa sẽ dẫn tới mùa đông dài không nắng. Chúng ta có thể sử dụng công nghệ hoặc để giải quyết tất cả những vấn đề này hoặc để tạo ra những vấn đề mới như biến đổi khí hậu, chiến tranh hạt nhân, đại dịch được thiết kế hoặc AI trở nên lệch lạc.

Tự huỷ diệt

Sau khi suy ngẫm những vấn đề mà công nghệ trong tương lai có thể gây ra, chúng ta cần cân nhắc thêm về những vấn đề mà việc không có công nghệ đưa tới. Với tinh thần này, chúng ta hãy khám phá viễn cảnh siêu trí tuệ không bao giờ được tạo ra bởi nhân loại tự tiêu diệt chính mình bằng cách này hay cách khác.

Chúng ta có thể làm điều đó bằng cách nào? Chiến thuật đơn giản nhất là “cứ đợi”. Mặc dù chúng ta sẽ thấy trong chương tiếp theo cách giải quyết các vấn đề như va đập với tiểu hành tinh và đại dương sôi sục, nhưng tất cả các giải pháp này đều đòi hỏi thứ công nghệ mà chúng ta vẫn chưa phát triển. Do vậy, trừ khi công nghệ tiến bộ vượt xa mức hiện tại, Mẹ Thiên nhiên sẽ khiến chúng ta tuyệt

chúng từ rất lâu trước cả khi một tỷ năm kịp trôi qua. Như nhà kinh tế học nổi tiếng John Maynard Keynes nói: "Trong tương lai tất cả chúng ta đều sẽ chết."

Thật không may, còn có những cách khiến chúng ta tự hủy diệt sớm hơn nhiều, đó là sự ngớ ngẩn tập thể. Tại sao giống loài chúng ta lại tự sát tập thể trong khi hầu như không ai muốn thế? Với trình độ hiện tại về trí tuệ và sự trưởng thành trong cảm xúc, con người vẫn có cái "khiếu" tính toán sai, hiểu lầm và thiếu năng lực, dẫn tới kết quả là lịch sử chúng ta tràn đầy tai nạn, chiến tranh và các thiên tai khác – những hậu quả mà, với nhận thức muộn màng, thực sự không ai muốn cả. Các nhà kinh tế học và toán học đã phát triển nên lý thuyết trò chơi thanh lịch, giải thích cho việc làm thế nào mà con người lại có thể bị lôi cuốn bởi những hành động gây ra kết quả thảm khốc cho tất cả mọi người.⁶

Chiến tranh hạt nhân:

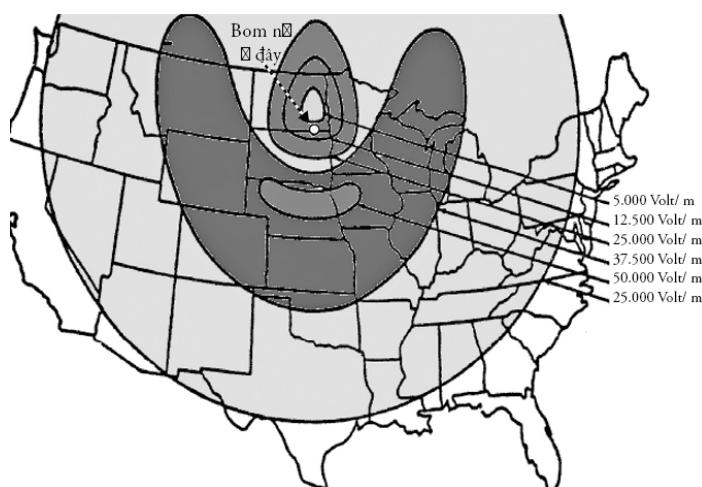
Tình huống thực tế về tính bất cẩn của loài người

Bạn nghĩ rằng nguy cơ càng cao thì chúng ta càng thận trọng hơn, nhưng việc xem xét kỹ hơn về rủi ro lớn nhất mà công nghệ hiện tại tạo ra là chiến tranh hạt nhân toàn cầu, cũng không đem lại cảm giác yên lòng hơn. Chúng ta vẫn đã và đang phải dựa vào may mắn để vượt qua một danh sách đáng xấu hổ những tai họa suýt xảy ra do đủ loại nguyên nhân: sự cố máy tính, mất điện, tình báo sai, lỗi điều hướng, tai nạn máy bay ném bom, nổ vệ tinh và vân vân.⁷ Trên thực tế, nếu không nhờ những hành động anh hùng của một số cá nhân – ví dụ, Vasili Arkhipov và Stanislav Petrov – một cuộc chiến hạt nhân toàn cầu có thể đã xảy ra. Với thực tế những gì đã xảy ra, nếu chúng ta tiếp tục hành động như hiện tại, tôi nghĩ rất khó có khả năng xác suất hàng năm của chiến tranh hạt nhân chỉ bằng một phần nghìn, hay xác suất xảy ra một cuộc chiến tranh trong vòng 10.000 năm sẽ vượt quá $1 - 0,99910000 \approx 99,995\%$.

Để đánh giá đầy đủ tính bất cẩn của con người, chúng ta phải nhận ra rằng mình đã bắt đầu canh bạc hạt nhân trước cả khi nghiên cứu kỹ các rủi ro. Thứ nhất, rủi ro bức xạ đã bị đánh giá thấp, và người ta đã phải chi trả hơn 2 tỷ đô-la để bồi thường cho các nạn

nhân phơi nhiễm bức xạ khi xử lý urani và nạn nhân của các thử nghiệm hạt nhân ở Hoa Kỳ nói riêng.⁸

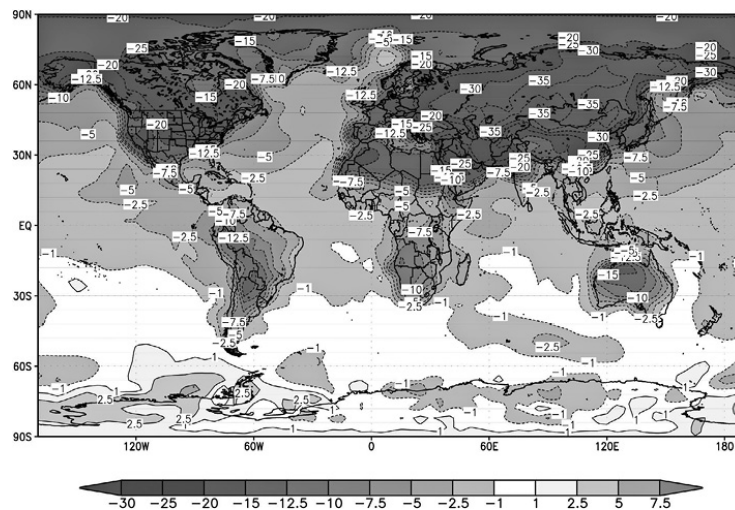
Thứ hai, cuối cùng người ta đã phát hiện ra rằng bom hiđrô được cho phát nổ hàng trăm kilômét bên trên Trái đất sẽ tạo ra xung điện từ mạnh (EMP) có khả năng vô hiệu hóa lưới điện và các thiết bị điện tử trên diện rộng (hình 5.2), khiến cơ sở hạ tầng bị tê liệt, các con đường bị tắc nghẽn do phương tiện không hoạt động và không có đủ các điều kiện lý tưởng để tồn tại hậu hạt nhân. Ví dụ, Ủy ban EMP Hoa Kỳ báo cáo rằng “cơ sở hạ tầng nước là một cỗ máy lớn, được hỗ trợ một phần bởi lực hấp dẫn nhưng chủ yếu vẫn là do điện” và việc thiếu nước có thể gây tử vong trong vòng ba tới bốn ngày.⁹



Hình 5.2: Chỉ một quả bom hiđrô phát nổ bên trên Trái đất 400 km cũng gây ra xung điện từ mạnh đủ để làm tê liệt các công nghệ dùng điện trên diện tích rộng. Bằng cách dịch chuyển điểm phát nổ sang phía đông nam, khu vực có hình như quả chuối vượt qua 37.500 volt trên mét có thể bao phủ phần lớn Đông duyên hải Hoa Kỳ. In lại từ Báo cáo Quân đội Hoa Kỳ AD- A278230 (không phải tài liệu mật), có thêm màu.

Thứ ba, phải đến bốn thập kỷ sau khi bắt đầu canh bạc thì người ra mới nhận ra khả năng xảy ra mùa đông hạt nhân, tức sau khi chúng ta đã thả 63.000 quả bom hiđrô! Bất kể thành phố nào bị cháy thì lượng khói khổng lồ lên tới tầng thượng lưu cũng có thể lan tỏa khắp thế giới, ngăn cản ánh sáng Mặt trời đủ để biến mùa hè thành mùa đông, giống như khi tiểu hành tinh hay siêu núi lửa gây ra tuyết

chúng hàng loạt trong quá khứ. Cảnh báo trong những năm 1980 mà các nhà khoa học của cả Mỹ và Liên Xô đưa ra về vấn đề này đã góp phần vào quyết định cắt giảm kho dự trữ của Ronald Reagan và Mikhail Gorbachev.¹⁰ Thật không may, các tính toán chính xác hơn đã vẽ nên bức tranh thậm chí còn ảm đạm hơn thế: hình 5.3 cho thấy nhiệt độ giảm khoảng 20°C (36°F) ở hầu hết các khu vực canh tác cốt lõi của Hoa Kỳ, châu Âu, Nga và Trung Quốc (và khoảng 35°C ở một số vùng của Nga) trong hai mùa hè đầu tiên, thậm chí vẫn tiếp tục giảm khoảng một nửa trong một thập kỷ sau đó.⁴⁸ Điều đó có nghĩa đơn giản là gì?



*Hình 5.3: Nhiệt độ giảm trung bình (theo °C) trong hai mùa hè đầu tiên sau cuộc chiến tranh hạt nhân diện rộng giữa Mỹ và Nga. Được in với sự đồng ý của Alan Robock.*¹¹

Ngay cả một người không có nhiều kinh nghiệm canh tác cũng có thể kết luận được rằng nhiệt độ mùa hè gần như đóng băng trong nhiều năm sẽ hủy diệt hầu hết quá trình sản xuất lương thực của chúng ta. Thật khó để dự đoán chính xác điều gì sẽ xảy ra sau khi hàng nghìn thành phố lớn nhất trên Trái đất trở thành đồng đổ nát và cơ sở hạ tầng toàn cầu sụp đổ, nhưng kể cả khi có một bộ phận dân số nhỏ nào đó không phải cúi đầu trước cái đói, sự giảm nhiệt độ hay bệnh dịch thì họ cũng sẽ cần phải đối phó với các băng nhóm vũ trang lang thang truy lùng thức ăn.

Tôi đã đi vào khá chi tiết về chiến tranh hạt nhân toàn cầu để quay lại trọng điểm rằng, không có nhà lãnh đạo thế giới “biết điều” nào mong muốn việc đó, nhưng nó có thể xảy ra do ngẫu nhiên. Điều này có nghĩa là chúng ta không thể tin tưởng những con người đồng hành của mình không bao giờ tự sát tập thể: việc chẳng ai muốn nó xảy ra không đồng nghĩa với việc nó sẽ được ngăn chặn để không xảy ra.

Thiết bị tận thế

Vậy con người có thể thực sự tự sát tập thể không? Ngay cả khi một cuộc chiến tranh hạt nhân toàn cầu có thể khiến 90% loài người thiệt mạng thì hầu hết các nhà khoa học vẫn dự đoán rằng tỷ lệ đó sẽ không tới mức 100% và do đó, chúng ta sẽ không rơi vào cảnh tuyệt chủng. Mặt khác, các câu chuyện về bức xạ hạt nhân, hạt nhân EMP và mùa đông hạt nhân đều chứng minh rằng mối nguy hiểm lớn nhất có lẽ lại đến từ những thứ mà chúng ta chưa từng nghĩ tới. Thật vô cùng khó khăn để dự đoán được về mọi khía cạnh sau thảm họa, cũng như về cách mà mùa đông hạt nhân, cơ sở hạ tầng sụp đổ, tỷ lệ đột biến tăng cao và bầu đàn vũ trang tuyệt vọng phản ứng trước các vấn đề khác như đại dịch mới, hệ sinh thái sụp đổ và những hệ quả mà chúng ta còn chưa từng tưởng tượng đến. Do đó, đánh giá cá nhân của tôi là mặc dù một chiến tranh hạt nhân xảy ra ngày mai khó có thể dẫn tới sự tuyệt chủng của con người nhưng chúng ta cũng không thể tự tin kết luận rằng xác suất này bằng không.

Khả năng xảy ra việc tự sát tập thể tăng lên nếu chúng ta nâng cấp vũ khí hạt nhân của ngày hôm nay thành một thiết bị tận thế có chủ đích. Được nhà chiến lược RAND Herman Kahn giới thiệu vào năm 1960 và phổ biến trong bộ phim Dr. Strangelove (Bác sĩ Strangelove) của Stanley Kubrick, thiết bị tận thế là hình mẫu về khả năng tận diệt. Đó là kẻ ngăn chặn hoàn hảo: một cỗ máy tự động trả thù mọi cuộc tấn công của đối phương bằng cách tàn sát tất cả nhân loại.

Một ứng cử viên cho thiết bị tận thế là kho lưu trữ ngầm khổng lồ của thứ mang tên vũ khí hạt nhân phóng xạ, có thể là những quả bom hiđrô khổng lồ bao quanh bởi lượng lớn coban. Nhà vật lý Leo Szilard đã tranh luận từ năm 1950 rằng việc này có thể giết chết tất cả con người trên Trái đất: các vụ nổ bom hiđrô sẽ khiến coban mang

tính phóng xạ và thổi nó vào tầng bình lưu, chu kỳ bán rã năm năm là đủ dài để nó lắng xuống khắp Trái đất (đặc biệt là nếu các thiết bị tận thế sinh đôi được đặt ở hai bán cầu đối diện), nhưng đủ cũng ngắn để gây ra cường độ bức xạ nguy hiểm. Báo cáo truyền thông cho thấy bom coban hiện đang được chế tạo lần đầu tiên. Khả năng xảy ra tự sát tập thể còn được củng cố nhờ việc thêm vào các quả bom được tối ưu hóa cho việc tạo ra mùa đông hạt nhân thông qua tối đa hóa lượng sol khí tồn tại lâu dài trong tầng bình lưu. Ưu thế chính của thiết bị tận thế là nó rẻ hơn nhiều so với chiến lược răn đe bằng hạt nhân thông thường: vì những quả bom này không cần bộ phóng nên chúng không cần hệ thống tên lửa đắt tiền, và bản thân việc chế tạo bom cũng rẻ hơn bởi nó không cần phải đủ nhẹ và rắn chắc để đặt vừa tên lửa.

Một khả năng khác là trong tương lai người ta sẽ khám phá ra thiết bị tận thế sinh học: một loại vi khuẩn hoặc virus được đặc chế để tiêu diệt tất cả con người. Nếu khả năng lây nhiễm của nó đủ cao và thời gian ủ bệnh đủ dài, về cơ bản mọi người đều có thể bị nhiễm bệnh trước cả khi họ nhận ra sự tồn tại của nó và có biện pháp đối phó. Ngay cả khi vũ khí sinh học kiểu này không thể huỷ diệt tất cả con người thì người ra cũng vẫn có một lý lẽ khác về mặt quân sự để ủng hộ việc chế tạo nó: thiết bị tận thế hiệu quả nhất là thiết bị kết hợp vũ khí hạt nhân, vũ khí sinh học và các loại khác nhằm tối đa hóa cơ hội ngăn chặn kẻ thù.

Vũ khí AI

Con đường công nghệ thứ ba dẫn tới tự sát tập thể có thể liên quan đến những vũ khí AI tương đối vô tri. Giả sử một siêu cường xây dựng hàng tỷ máy bay tấn công không người lái có kích cỡ ong đất như ở chương 3 và sử dụng chúng để giết bất cứ ai ngoại trừ công dân và đồng minh của họ, sau khi xác định danh tính những người này từ xa qua thẻ ID có tần số vô tuyến giống như hầu hết các sản phẩm siêu thị hiện nay. Các thẻ này có thể được phát cho tất cả các công dân để đeo trên vòng tay hoặc cấy vào dưới da, như trong phần chế độ chuyên chế. Điều này có lẽ sẽ thúc đẩy một siêu cường đối lập xây dựng thứ gì đó tương tự. Khi chiến tranh vô tình nổ ra, tất cả con người sẽ bị tiêu diệt, kể cả những bộ tộc xa xôi không liên quan, bởi sẽ không có ai đeo cả hai loại thẻ ID này. Kết hợp điều này

với thiết bị tận thế hạt nhân và sinh học sẽ đẩy cơ hội tự sát tập thể thành công lên cao hơn nữa.

Bạn muốn gì?

Khi bắt đầu chương này, bạn cân nhắc mình muốn cuộc đua AGI hiện tại dẫn tới đâu. Giờ chúng ta đã cùng nhau khám phá một loạt các viễn cảnh, vậy viễn cảnh nào thu hút bạn và bạn nghĩ chúng ta nên cố gắng hết sức để tránh xa viễn cảnh nào? Bạn có thích viễn cảnh nào hơn hẳn không? Hãy cho tôi và bạn đọc biết tại <http://AgeOfAi.org>, và cùng tham gia thảo luận!

Rõ ràng chúng ta không nên xem các viễn cảnh đã xem xét là một danh sách đầy đủ, thậm chí nhiều viễn cảnh trong đó còn thiếu nhiều chi tiết, nhưng tôi đã cố gắng hết sức để bao quát và tổng hợp toàn bộ các khả năng từ công nghệ cao đến công nghệ thấp tới không có công nghệ, đồng thời mô tả tất cả những hy vọng và nỗi sợ chính yếu được thể hiện trong văn học.

Một trong những phần thú vị nhất khi viết cuốn sách này là nghe xem bạn bè và đồng nghiệp tôi nghĩ gì về những viễn cảnh này, và tôi đã rất thích thú khi nhận thấy chẳng có sự thống nhất nào cả. Điều duy nhất tất cả mọi người đồng ý là các tình huống này đều tinh vi hơn vẻ ban đầu. Những người thích một viễn cảnh nào đó hơn thường cũng đồng thời nhận thấy một số khía cạnh gây khó chịu của nó. Đối với tôi, điều này có nghĩa là con người chúng ta cần tiếp tục, đào sâu thêm những trao đổi về các mục tiêu tương lai này, để chúng ta biết nên cầm lái theo hướng nào. Tiềm năng tương lai cho cuộc sống vũ trụ của chúng ta là vô cùng vô cùng lớn, vì vậy đừng bỏ qua nó như một con tàu trôi dạt không đích đến, không biết nơi chúng ta muốn đi là đâu!

Tiềm năng trong tương lai này thật sự lớn tới đâu? Bất kể công nghệ của chúng ta tiến bộ như thế nào, khả năng cải thiện và lan truyền của Sự sống 3.0 vào vũ trụ của chúng ta sẽ bị giới hạn bởi các định luật vật lý – những giới hạn tối hậu này là gì trong hàng tỷ năm tới? Liệu lúc này Vũ trụ của chúng ta đã đầy ắp sự sống ngoài hành tinh chưa hay chúng ta vẫn đơn độc? Điều gì sẽ xảy ra nếu các nền

văn minh vũ trụ đang phát triển gặp gỡ nhau? Chúng ta sẽ giải quyết những câu hỏi hấp dẫn này trong chương tiếp theo.

TÓM TẮT:

- Cuộc đua hiện tại tới AGI có thể kết thúc theo một loạt các viễn cảnh hệ quả hấp dẫn trong thiên niên kỷ sắp tới.
- Siêu trí tuệ có thể cùng tồn tại hòa bình với con người hoặc bởi nó buộc phải vậy (viễn cảnh vị thần bị nô dịch) hoặc vì nó là “AI thân thiện” mong muốn vậy (các viễn cảnh xã hội tự do lý tưởng, thần bảo hộ, nhà độc tài nhân từ và bảo vệ vườn thú).
- Siêu trí tuệ có thể bị cản trở bởi một AI (viễn cảnh người gác cửa) hoặc do con người (viễn cảnh 1984), bằng cách cố tình quên đi công nghệ siêu trí tuệ (viễn cảnh nghịch chuyển) hoặc do thiếu động lực xây dựng nó (viễn cảnh xã hội bình đẳng lý tưởng).
- Nhân loại có thể tuyệt chủng và bị thay thế bởi AI (viễn cảnh kẻ chinh phục và hậu duệ) hoặc bởi không gì cả (viễn cảnh tự hủy diệt).
- Hoàn toàn không có sự đồng thuận về việc viễn cảnh nào trong số này là đáng được mong đợi, và tất cả đều chứa đựng các yếu tố bị phản đối. Điều này khiến việc tiếp tục, đào sâu hơn cuộc đối thoại về mục tiêu tương lai càng quan trọng hơn, để từ đó chúng ta không vô tình trôi dạt hoặc bẻ lái theo hướng đi không may mắn.

Chương 6

Tài nguyên vũ trụ: Một tỷ năm sau và xa hơn nữa

Suy đoán của chúng tôi dừng lại ở siêu văn minh, nơi tổng hòa tất cả sự sống trên hệ Mặt trời, liên tục cải tiến và mở rộng chính nó, từ Mặt trời lan tỏa ra xa, biến đổi thứ vô tri thành trí óc.

Hans Moravec, Những đứa con trí óc

Đối với tôi, phát kiến khoa học tạo cảm hứng nhất cho đến nay là chúng ta đã cực kỳ đánh giá thấp tiềm năng tương lai của sự sống. Những ước mơ và khát vọng không cần giới hạn tuổi thọ trong chưa đầy một thế kỷ, suy mòn bởi bệnh tật, đói nghèo và sự hoang mang. Thay vào đó, với trợ giúp của công nghệ, sự sống có tiềm năng nở rộ hàng tỷ năm, không chỉ ở đây, trong Hệ Mặt Trời, mà còn ra khắp vũ trụ rộng lớn hơn và hừng khởi hơn nhiều so với những gì tổ tiên ta tưởng tượng. Ngay cả trời xanh cũng không còn là giới hạn nữa.

Đây là tin đáng mừng đối với một loài vốn luôn có cảm hứng từ việc đẩy lùi các giới hạn trong suốt các thời đại. Thế vận hội Olympic tôn vinh việc đẩy lùi các giới hạn về sức mạnh, tốc độ, sự nhanh nhẹn khéo léo và sức bền. Khoa học tôn vinh việc đẩy lùi giới hạn của tri thức và hiểu biết. Văn học nghệ thuật tôn vinh việc đẩy lùi giới hạn của sáng tạo cái đẹp hay những trải nghiệm làm phong phú cho đời. Nhiều người, nhiều tổ chức và quốc gia ăn mừng việc gia tăng nguồn tài nguyên, lãnh thổ và tuổi thọ. Với sự ám ảnh vốn có của con người về những giới hạn, cũng dễ hiểu khi cuốn sách bán chạy nhất mọi thời đại chính là Sách Kỷ lục Thế giới Guinness.

Vậy nếu công nghệ có thể phá vỡ giới hạn của sự sống mà ta vẫn thấy thì giới hạn tận cùng là gì? Bao nhiêu phần trong vũ trụ có thể sống dậy? Sự sống có thể tiến xa đến đâu và kéo dài đến bao giờ? Sự sống có thể sử dụng bao nhiêu vật chất, hay có thể xuất ra được bao nhiêu năng lượng, thông tin và tính toán? Những giới hạn tận cùng này được đặt ra không phải bởi hiểu biết của chúng ta, mà bởi các định luật vật lý. Trớ trêu thay, điều này khiến việc phân tích tương lai dài hạn của sự sống lại dễ dàng hơn tương lai ngắn hạn.

Nếu lịch sử vũ trụ 13,8 tỷ năm được thu ngắn lại trong một tuần, thì câu chuyện ly kỳ kéo dài 10000 năm trong hai chương vừa rồi sẽ trôi qua trong chưa đầy nửa giây. Nghĩ là mặc dù ta không thể dự đoán liệu bùng nổ trí tuệ có khả năng diễn ra không, diễn ra như thế nào và kết quả ngay sau đó sẽ ra sao, nhưng tất cả sự hỗn loạn này chỉ là một tích tắc trong lịch sử vũ trụ và các chi tiết trong đó không ảnh hưởng đến giới hạn tận cùng của sự sống. Nếu sự sống sau bùng nổ cũng ám ảnh như con người ngày nay về việc đẩy lùi giới hạn, thì nó sẽ phát triển công nghệ để thật sự chạm đến các giới hạn đó – vì nó có khả năng đó. Trong chương này, ta sẽ tìm hiểu những giới hạn này là gì, từ đó có được cái nhìn về tương lai lâu dài của sự sống. Vì những giới hạn này dựa trên hiểu biết hiện tại của ta về vật lý học, chúng nên được coi như mức khả năng “sàn”: các phát kiến khoa học trong tương lai có thể đem lại cơ hội làm tốt hơn thế.

Nhưng ta có biết chắc rằng sự sống tương lai sẽ đầy tham vọng như vậy không? Không. Có thể sự sống sẽ trở nên an phận như người nghiện ma túy hay nghiện ti-vi, chỉ ngồi xem không ngừng các phần chiếu lại của chương trình Theo chân nhà Kardashian. Tuy nhiên, chúng ta có lý do để cho rằng tham vọng là một đặc tính khá phổ thông của sự sống bậc cao. Gần như bất kể nó đang cố gắng tối đa hóa điều gì, dù đó là trí tuệ, tuổi thọ, kiến thức hay trải nghiệm thú vị, thì sự sống bậc cao cũng sẽ cần tài nguyên. Do đó nó có động cơ để thúc đẩy công nghệ đến giới hạn tận cùng, để tận dụng tối đa tài nguyên mà nó có. Sau đó, cách duy nhất để tiếp tục cải tiến là thu thập nhiều tài nguyên hơn, bằng cách mở rộng ra những vùng ngày càng rộng lớn của vũ trụ.

Ngoài ra, sự sống có thể độc lập sinh ra từ nhiều nơi trong vũ trụ. Trong trường hợp đó, các nền văn minh không tham vọng đơn giản là trở nên không còn giá trị liên quan tới vũ trụ nữa, bởi những dạng sống nhiều tham vọng nhất cuối cùng sẽ lấy đi các phần ngày càng lớn của tài nguyên vũ trụ. Chọn lọc tự nhiên do đó diễn ra ở quy mô toàn vũ trụ, và sau một thời gian, gần như tất cả sự sống tồn tại đều là sự sống có tham vọng. Tóm lại, nếu ta quan tâm tới vấn đề bao nhiêu phần của vũ trụ có thể sống dậy, ta nên nghiên cứu những giới hạn của tham vọng mà các định luật vật lý đặt ra. Vậy thì làm thôi! Đầu tiên hãy cùng khám phá giới hạn của những gì ta có thể làm

được với các tài nguyên (vật chất, năng lượng, v.v.) mà ta có trong Hệ Mặt Trời, sau đó tìm hiểu cách để kiếm được nhiều tài nguyên hơn thông qua khám phá và định cư trong vũ trụ.

Tận dụng tối đa nguồn tài nguyên

Trong khi các siêu thị và điểm giao dịch hàng hóa bán hàng vạn sản phẩm mà ta có thể gọi là “tài nguyên” thì sự sống trong tương lai, khi đã đạt đến giới hạn công nghệ, sẽ chỉ cần chủ yếu một tài nguyên cơ bản gọi là vật chất hạt cơ bản, nghĩa là bất kỳ thứ gì làm từ các nguyên tử hay các hạt cấu tạo nên nguyên tử (hạt quark và electron). Bất kể vật chất này ở dạng nào, công nghệ bậc cao cũng có thể sắp xếp lại nó thành bất kỳ chất hay vật thể mong muốn nào, gồm cả nhà máy điện, máy tính và các dạng sống bậc cao. Vì vậy, hãy bắt đầu bằng các xem xét các giới hạn của loại năng lượng cung cấp sức mạnh cho sự sống bậc cao và quá trình xử lý thông tin giúp nó có thể tư duy.

Xây dựng quả cầu Dyson

Khi nói đến tương lai sự sống, một trong những người có tầm nhìn đầy hy vọng nhất là Freeman Dyson. Tôi có vinh dự được biết ông trong hai thập kỷ qua, nhưng khi tôi gặp ông lần đầu, tôi vẫn rất hồi hộp. Khi ấy tôi còn là nghiên cứu sinh sau tiến sĩ, đang ngồi nhai cơm với bạn trong phòng ăn trưa ở Viện Nghiên cứu Cao cấp Princeton thì như từ trên trời rơi xuống, nhà vật lý nổi tiếng thế giới đã từng trò chuyện với Einstein và Gödel đi tới và tự giới thiệu mình, hỏi xem ông có thể ngồi cùng chúng tôi không! Tuy nhiên, ông nhanh chóng khiến tôi an tâm, bằng cách giải thích là ông thích ăn trưa cùng các bạn trẻ hơn là những giáo sư già buồn tẻ. Mặc dù khi tôi gõ những dòng này thì ông đã 93 tuổi, nhưng tinh thần Freeman vẫn trẻ trung hơn hầu hết những người tôi biết, và ánh nhìn tinh nghịch trong mắt ông cho thấy ông chẳng hề quan tâm đến hình thức, thứ bậc học thuật hay những quan niệm truyền thống. Ý tưởng càng táo bạo càng khiến ông hứng thú.

Khi chúng tôi nói chuyện về sử dụng năng lượng, ông chế nhạo con người chúng ta là thiếu tham vọng, chỉ ra rằng ta có thể đáp ứng nhu cầu năng lượng ngày nay trên khắp toàn cầu bằng các thu giữ

ánh Mặt Trời chiếu sáng trong một diện tích nhỏ hơn 0,5% của sa mạc Sahara. Nhưng sao phải dừng lại ở đó? Tại sao lại dừng lại ở việc thu giữ tất cả ánh sáng Mặt Trời chiếu lên Trái Đất và để cho hầu hết ánh sáng đó tỏa ra không gian một cách phí phạm? Sao không đơn giản là đưa tất cả năng lượng của Mặt Trời sản xuất ra vào việc phục vụ cho cuộc sống?

Lấy cảm hứng từ truyện khoa học viễn tưởng kinh điển năm 1937 *Star Maker* (Người làm ra ngôi sao) của Olaf Stapledon, trong đó có những vành đai gồm các thế giới nhân tạo xoay quanh quỹ đạo của sao mẹ, Freeman Dyson xuất bản một bài miêu tả vào năm 1960 về thứ sau này được biết đến với tên gọi quả cầu Dyson.¹ Ý tưởng của Freeman là sắp xếp lại Sao Mộc thành một sinh cầu ở dạng lớp vỏ hình cầu bao quanh Mặt Trời, nơi con cháu chúng ta có thể sinh sôi, tận hưởng sinh khối lớn gấp 100 tỷ lần và nguồn năng lượng gấp 1000 tỷ lần nhân loại sử dụng ngày nay.² Ông lập luận rằng đây là bước phát triển tự nhiên tiếp theo: “Người ta nên kỳ vọng rằng, trong vòng một vài ngàn năm sau khi bước vào giai đoạn phát triển công nghiệp, bất kỳ loài thông minh nào cũng có thể chiếm đóng một sinh cầu nhân tạo bao phủ hoàn toàn sao mẹ.” Cuộc sống ở phía bên trong quả cầu Dyson sẽ không có buổi đêm: bạn sẽ luôn nhìn thấy Mặt Trời ngay trên đầu mình và khắp bầu trời sẽ chỉ thấy ánh nắng Mặt Trời phản chiếu từ phần còn lại của sinh cầu, giống như ngày nay bạn có thể thấy ánh sáng phản chiếu từ Mặt Trăng vào ban ngày. Nếu muốn ngắm sao, bạn chỉ cần đi “lên gác” và nhìn ra ngoài vũ trụ từ phía bên ngoài quả cầu Dyson.

Một cách làm công nghệ thấp để xây dựng một phần quả cầu Dyson là đặt một vành đai các nơi cư trú theo vòng tròn xung quanh Mặt Trời. Để bao phủ hoàn toàn Mặt Trời, bạn có thể thêm các vành đai quay quanh Mặt Trời theo các trục khác nhau ở khoảng cách khác biệt đôi chút để tránh va chạm. Nhằm tránh mỗi phiền phức là những vành đai di chuyển nhanh này không thể kết nối với nhau, vì thế gây rắc rối cho giao thông và liên lạc, người ta cũng có thể xây dựng một quả cầu Dyson nguyên khối đứng yên ở nơi lực hút trọng lực của Mặt Trời được cân bằng bởi lực đẩy của bức xạ Mặt Trời – một ý tưởng do Robert L. Forward và Colin McInnes đề xướng. Quả

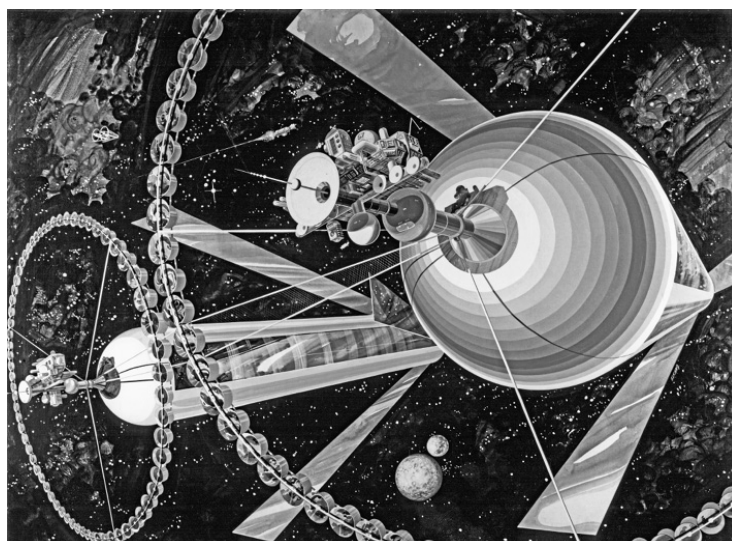
cầu đó có thể được xây dựng bằng cách dần dần thêm vào nhiều “vệ tinh, tức vệ tinh nhân tạo tĩnh, đối kháng với trọng lực của Mặt Trời bằng áp lực bức xạ thay vì lực ly tâm. Cả hai lực này giảm đi theo bình phương khoảng cách tính từ Mặt Trời, nghĩa là nếu có thể cân bằng chúng tại một khoảng cách nào đó với Mặt Trời thì chúng cũng sẽ cân bằng ở bất kỳ khoảng cách nào, tạo điều kiện thuận lợi cho việc tự do dừng đỗ ở bất kỳ đâu trong Hệ Mặt Trời. Vệ tinh cần là những tấm cực kỳ nhẹ, chỉ nặng khoảng 0,77 gram trên mỗi mét vuông, tức là nhẹ hơn giấy khoảng 100 lần. Nhưng vấn đề khối lượng này cũng khó có thể là một yếu tố gây cản trở. Ví dụ, một tấm graphene (một lớp nguyên tử cacbon xếp thành các hình lục giác trông như lưới mắt cáo) còn nhẹ hơn giới hạn đó tới cả ngàn lần. Nếu quả cầu Dyson được xây dựng để phản chiếu thay vì hấp thụ hầu hết ánh sáng Mặt Trời, thì tổng cường độ của ánh sáng phản chiếu xung quanh bên trong nó sẽ được tăng lên đáng kể, càng làm tăng lượng áp lực bức xạ và khối lượng có thể nâng đỡ được bên trong quả cầu. Nhiều ngôi sao khác có độ sáng gấp ngàn lần, thậm chí triệu lần hơn Mặt Trời, do đó có thể nâng đỡ những quả cầu Dyson đứng yên với khối lượng lớn hơn tương ứng.

Nếu ta muốn có một quả cầu Dyson rắn chắc nặng hơn nhiều ở đây, ngay trong Hệ Mặt Trời, thì việc chống lại trọng lực của Mặt trời sẽ cần một loại vật liệu cực mạnh có thể chịu được áp lực gấp hàng vạn lần áp lực ở chân các tòa nhà chọc trời cao nhất thế giới mà không tan chảy hay oằn xuống. Để có thể thọ lâu, một quả cầu Dyson cần phải năng động và thông minh, liên tục điều chỉnh vị trí và hình dáng của mình, sao cho những thiên thạch và sao chổi phiến phức sẽ bay ngang qua mà không gây ảnh hưởng. Hoặc người ta có thể sử dụng một hệ thống phát hiện và làm chệch hướng để xử lý những kẻ xâm phạm hệ thống, công phá chúng thành nhiều mảnh nếu muốn và dùng vật chất đó vào việc có ích.

Với loài người ngày nay, cuộc sống ở trên hay ở trong quả cầu Dyson nếu may mắn thì có vẻ mất phương hướng còn xui xẻo thì bất khả thi, nhưng điều đó không chắc sẽ ngăn cản các dạng sống sinh học hoặc phi sinh học trong tương lai sống ở đó. Nếu quả cầu đó chuyển động quanh quỹ đạo, nó sẽ đem lại môi trường gần như không trọng lực, còn nếu đó là một quả cầu đứng yên, bạn sẽ chỉ có

thể đi lại ở mặt phía ngoài (hướng ra xa khỏi Mặt Trời) mà không bị rơi ra, với trọng lực yếu hơn khoảng mười ngàn lần so với trọng lực bạn vẫn quen. Bạn sẽ không có từ trường (trừ khi bạn làm ra một cái) để bảo vệ bạn khỏi những hạt nguy hại từ Mặt Trời. Điều an ủi là một quả cầu Dyson với kích cỡ quỹ đạo hiện tại của Trái Đất sẽ đem lại cho chúng ta diện tích bề mặt lớn hơn khoảng 500 triệu lần để sống.

Nếu ta muốn có những nơi cư ngụ giống Trái Đất hơn thì tin vui là chúng dễ thực hiện hơn quả cầu Dyson. Ví dụ, hình 6.1 và 6.2 cho thấy thiết kế về một môi trường nhân tạo hình trụ do nhà vật lý người Mỹ Gerard K. O'Neill khởi xướng, có các tính năng trọng lực nhân tạo, che chắn các tia vũ trụ, chu kỳ ngày đêm 24 giờ, cùng bầu khí quyển và hệ sinh thái như Trái Đất.



Hình 6.1: Một cặp hình trụ O'Neill xoay ngược chiều nhau có thể đem lại nơi ở như Trái Đất cho con người nếu chúng quay quanh Mặt Trời và luôn hướng thẳng vào Mặt Trời. Lực ly tâm từ việc xoay quanh trục đem lại trọng lực nhân tạo, và ba tấm gương gấp lại được sẽ phản chiếu ánh Mặt Trời vào trong theo chu kỳ ngày đêm 24 giờ. Những nơi ở nhỏ hơn sắp xếp thành vòng tròn và được chuyên biệt hóa cho nông nghiệp. Hình ảnh do Rick Guidice/NASA cung cấp.

Những nơi ở như vậy có thể tự do quay quanh quỹ đạo bên trong một quả cầu Dyson, hoặc người ta có thể chỉnh sửa ra những dạng nơi ở tương tự để gắn liền bên ngoài quả cầu này.

Xây dựng các nhà máy điện tốt hơn

Mặc dù quả cầu Dyson có hiệu quả năng lượng cao theo tiêu chuẩn kỹ thuật ngày nay, nhưng chúng vẫn còn xa mới đạt tới mức đẩy lùi được những giới hạn đặt ra bởi các định luật vật lý. Einstein đã cho thấy nếu ta có thể biến đổi khối lượng thành năng lượng với hiệu suất 100%⁴⁹ thì một khối lượng

m

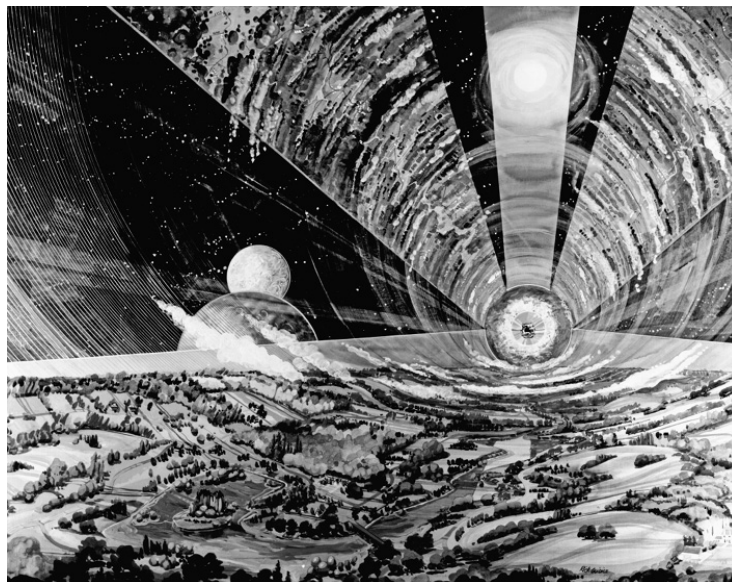
sẽ cho ta một lượng năng lượng

E

theo công thức nổi tiếng của ông

$$E = mc^2$$

, trong đó c là tốc độ ánh sáng. Nghĩa là, vì c là con số rất lớn, nên một khối lượng nhỏ cũng có thể tạo ra một lượng năng lượng khổng lồ.



Hình 6.2: Nhìn từ bên trong một trong hai hình trụ O'Neill ở hình trước. Nếu nó có đường kính là 6,4 km và cứ xoay quanh trục mỗi hai phút một lần, người trên bề mặt hình trụ sẽ có cảm giác trọng lực

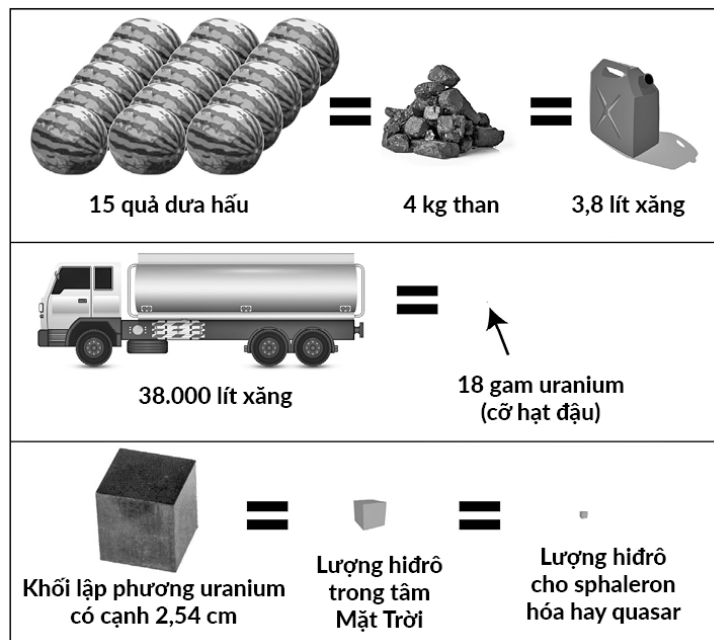
như ở Trái Đất. Mặt Trời ở phía sau nhưng lại trông như ở trên đầu, nhờ một tấm gương bên ngoài hình trụ cứ đến đêm thì gấp lại. Các cửa sổ khép kín giữ cho không khí không thoát ra ngoài hình trụ. Hình ảnh do Rick Guidice/NASA cung cấp.

Nếu ta có một nguồn cung phản vật chất dồi dào (mà hiện ta không có), thì việc tạo ra một nhà máy điện hiệu suất 100% sẽ rất dễ dàng: chỉ cần rót một thìa cà phê phản-nước vào nước bình thường là sẽ cho ra năng lượng tương đương với 200.000 tấn thuốc nổ TNT hay sức mạnh của một quả bom hiđrô – đủ để cung cấp cho nhu cầu điện năng của cả thế giới trong khoảng bảy phút.

Nhưng trong thực tế, các cách tạo ra năng lượng thường gặp nhất ngày nay lại thiếu hiệu quả đến đau lòng, như tóm tắt trong bảng 6.1 và hình 6.3. Việc tiêu hóa một thanh kẹo sô-cô-la có hiệu suất chỉ 0,00000001%, nghĩa là nó tạo ra chỉ một phần một nghìn tỷ năng lượng mc^2 mà nó chứa. Chỉ cần dạ dày bạn có hiệu suất được 0.001% thôi là bạn đã có thể ăn một bữa là đủ no cả đời. So với việc ăn uống, việc đốt than và xăng chỉ hiệu quả hơn lần lượt ba và năm lần.

<i>Phương pháp</i>	<i>Hiệu suất</i>
<i>Tiêu hóa thanh kẹo sô-cô-la</i>	<i>0,00000001%</i>
<i>Đốt cháy than</i>	<i>0,00000003%</i>
<i>Đốt cháy xăng</i>	<i>0,00000005%</i>
<i>Phân rã nguyên tử uranium-235</i>	<i>0,08%</i>
<i>Sử dụng quả cầu Dyson đến khi Mặt Trời tắt</i>	<i>0,08%</i>
<i>Tổng hợp hạt nhân hiđrô với heli</i>	<i>0,7%</i>
<i>Quay tròn lỗ đen</i>	<i>29%</i>
<i>Quả cầu Dyson quanh quasar</i>	<i>42%</i>
<i>Sphaleron hóa</i>	<i>50%?</i>
<i>Sự bốc hơi lỗ đen</i>	<i>90%</i>

Bảng 6.1: Hiệu suất của việc chuyển đổi khối lượng thành năng lượng có ích theo giới hạn lý thuyết $E = mc^2$. Như giải thích trong chương này, đạt được hiệu suất 90% từ việc nuôi các lỗ đen và đợi chúng bốc hơi không may là quá chậm nên không thể sử dụng, và làm tăng tốc quá trình này thì lại giảm hiệu suất đi đáng kể.



Hình 6.3: Công nghệ tiên tiến có thể chiết xuất từ vật chất lượng năng lượng nhiều hơn đáng kể so với khi ta ăn hay đốt cháy vật chất, và ngay cả phản ứng tổng hợp hạt nhân (phản ứng nhiệt hạch) cũng chiết xuất ít hơn 140 lần so với mức giới hạn năng lượng mà các quy luật vật lý đặt ra. Các nhà máy điện khai thác sphaleron, quasar (vật thể giống sao) hay lỗ đen bốc hơi có thể hiệu quả hơn nhiều.

Các lò phản ứng hạt nhân ngày nay hoạt động hiệu quả hơn nhiều, bằng cách chia nhỏ nguyên tử uranium thông qua phản ứng phân rã nguyên tử (phản ứng phân hạch), nhưng vẫn không chiết xuất được quá 0,08% năng lượng của các nguyên tử này. Phản ứng hạt nhân trong tâm Mặt Trời có hiệu suất bằng bình phương của những lò phản ứng do con người xây dựng, chiết xuất được 0,7% năng lượng từ hiđrô bằng cách tổng hợp hiđrô thành heli. Tuy nhiên, dù có bao bọc Mặt Trời trong một quả cầu Dyson hoàn hảo thì ta cũng sẽ không bao giờ chuyển hóa được hơn 0,08% khối lượng của Mặt Trời thành năng lượng có ích, vì một khi đã tiêu thụ hết khoảng một phần mười nhiên liệu hiđrô của mình, Mặt Trời sẽ chấm dứt vòng đời của một ngôi sao bình thường, phình to thành một sao khổng lồ đỏ rồi bắt đầu chết dần. Mọi chuyện cũng không khá hơn cho các ngôi sao khác: tỷ lệ tiêu thụ hiđrô trong vòng đời chính là khoảng từ 4% đối với sao rất nhỏ đến 12% đối với những sao lớn

nhất. Kể cả khi hoàn thiện được một lò phản ứng nhiệt hạch nhân tạo có thể tổng hợp 100% lượng nguyên tử hiđrô mà ta có trong tay, thì ta vẫn sẽ mắc kẹt ở hiệu suất thấp đến xấp xỉ 0,7% của toàn bộ quá trình nhiệt hạch. Làm thế nào để hiệu quả hơn?

Bốc hơi lỗ đen

Trong cuốn

Lược sử thời gian,

Stephen Hawking đề xuất một nhà máy điện lỗ đen.⁵⁰ Điều này nghe như một nghịch lý vì từ lâu người ta vẫn tin lỗ đen là những cái bẫy mà không một thứ gì, kể cả ánh sáng, có thể thoát khỏi. Tuy nhiên, Hawking nổi tiếng vì đã tính toán ra rằng hiệu ứng trọng lực lượng tử khiến lỗ đen phản ứng như một vật thể nóng – càng nhỏ, càng nóng – tỏa ra bức xạ nhiệt hiện được gọi là

bức xạ Hawking.

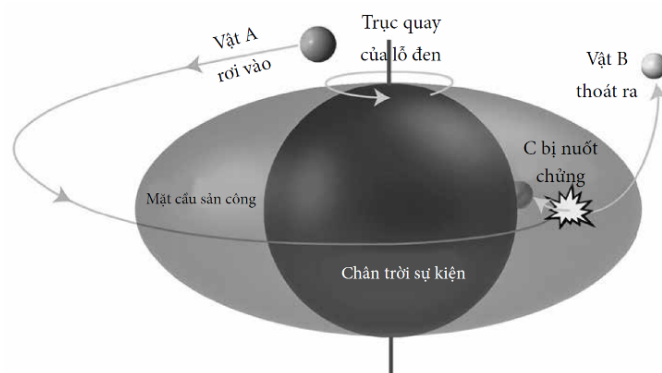
Điều này nghĩa là lỗ đen dần dần mất năng lượng và bốc hơi đi mất. Nói cách khác, bạn thả bất cứ vật chất gì vào lỗ đen thì vật chất đó cuối cùng cũng sẽ trở lại ở dạng bức xạ nhiệt, nên cho đến khi lỗ đen bốc hơi hoàn toàn, bạn đã chuyển hóa vật chất thành bức xạ với hiệu suất gần 100%.⁵¹

Việc sử dụng sự bốc hơi lỗ đen như một nguồn năng lượng có một vấn đề là trừ khi lỗ đen đó có kích cỡ nhỏ hơn nhiều so với một nguyên tử, quá trình bốc hơi đó sẽ cực kỳ chậm chạp, tới mức phải mất lâu hơn cả tuổi đời hiện tại của Vũ trụ mới bức xạ được lượng năng lượng ít hơn cả một ngọn nến. Điện năng tạo ra giảm dần theo bình phương kích thước lỗ đen, do đó các nhà vật lý Louis Crane và Shawn Westmoreland đã đề xuất sử dụng lỗ đen nhỏ hơn proton khoảng một nghìn lần và nặng cỡ tàu biển lớn nhất từng ra khơi.³ Động lực chính của họ là dùng động cơ lỗ đen đó để chạy một con tàu vũ trụ (ta sẽ bàn đến chủ đề này ở dưới đây), nên họ quan tâm tới khả năng di động nhiều hơn là hiệu suất và đề xuất nuôi lỗ đen bằng ánh sáng laser, hoàn toàn không gây ra chuyển hóa theo công thức năng lượng-vật chất. Dù bạn có thể nuôi lỗ đen bằng vật chất

thay vì bức xạ đi chẳng nữa thì việc đảm bảo hiệu suất cao vẫn có vẻ khó khăn: để khiến các proton đi vào một lỗ đen nhỏ hơn chúng một nghìn lần, chúng phải được bắn vào lỗ đen đó với một chiếc máy có sức mạnh ngang với máy gia tốc hạt lớn (Large Hadron Collider hay LHC), gia tăng năng lượng mc^2 của chúng với động năng lớn gấp ít nhất cả nghìn lần. Vì ít nhất 10% động năng đó sẽ bị mất cho các hạt graviton khi lỗ đen bốc hơi, nên lượng năng lượng ta đưa vào lỗ đen sẽ lớn hơn ta có thể lấy ra và sử dụng, kết quả là hiệu suất bằng âm. Yếu tố khác gây khó khăn cho viễn cảnh nhà máy điện lỗ đen là ta vẫn thiếu một lý thuyết vững chắc về trọng lực lượng tử để dựa vào đó thực hiện các tính toán – nhưng sự thiếu chắc chắn này dĩ nhiên cũng có thể đồng nghĩa với việc tồn tại những hiệu ứng trọng lực lượng tử mới và hữu dụng mà ta chưa khám phá ra.

Quay tròn lỗ đen

May thay, có những cách khác để sử dụng lỗ đen làm nhà máy điện mà không liên quan đến trọng lực lượng tử hay những lý thuyết vật lý chưa nắm rõ khác. Ví dụ, nhiều lỗ đen đang quay tròn rất nhanh với chân trời sự kiện xoay vòng vòng gần bằng tốc độ ánh sáng, và năng lượng quay tròn này có thể chiết xuất được. Chân trời sự kiện của lỗ đen là khu vực nơi ngay cả ánh sáng cũng không thể thoát ra ngoài, vì lực hấp dẫn quá mạnh. Hình 6.4 minh họa bên ngoài chân trời sự kiện, một lỗ đen quay tròn còn có một khu vực gọi là mặt cầu sản công, nơi mà lỗ đen quay tròn kéo theo không gian chuyển động cùng nó nhanh đến nỗi không thể có hạt nào đứng yên mà không chuyển động theo.



Hình 6.4: Một phần năng lượng quay tròn của lỗ đen quay tròn có thể chiết xuất bằng cách ném vật A vào gần lỗ đen và để nó tách đôi thành hai phần, phần C bị nuốt chửng còn phần B thoát ra với năng lượng lớn hơn A có ban đầu.

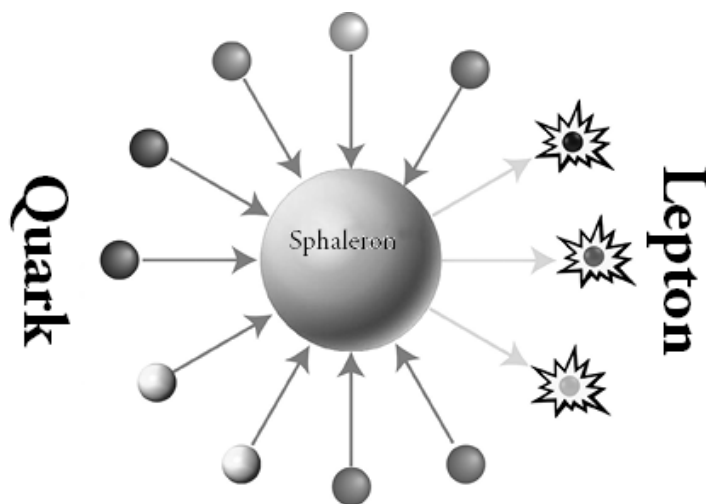
Nếu bạn ném một vật thể vào trong mặt cầu sản công, nó sẽ tăng tốc khi xoay tròn quanh lỗ đen. Không may là nó sẽ nhanh chóng bị lỗ đen nuốt chửng, mãi mãi biến mất phía sau chân trời sự kiện, vậy nên việc thử chiết xuất năng lượng không mang lại kết quả. Tuy nhiên, Roger Penrose đã phát hiện rằng nếu bạn phóng vật thể vào ở một góc khéo léo và khiến nó tách đôi thành hai mảnh như hình 6.4 minh họa, thì bạn có thể sắp xếp sao cho chỉ một mảnh bị nuốt chửng còn mảnh kia thoát ra khỏi lỗ đen với năng lượng nhiều hơn năng lượng khi bạn phóng đi. Nói cách khác, bạn đã chuyển hóa thành công một phần năng lượng quay tròn của lỗ đen thành năng lượng có ích mà bạn có thể khai thác. Bằng cách lặp lại quá trình này nhiều lần, bạn có thể vắt kiệt toàn bộ năng lượng quay tròn của lỗ đen, khiến nó ngừng quay và mặt cầu sản công của nó biến mất. Nếu tốc độ quay tròn của lỗ đen ban đầu đạt mức như tự nhiên cho phép, với chân trời sự kiện chuyển động với tốc độ tương đương ánh sáng, thì chiến thuật này sẽ cho phép bạn chuyển hóa 29% khối lượng lỗ đen thành năng lượng. Vẫn còn nhiều điều chưa chắc chắn về việc những lỗ đen trên bầu trời đêm của chúng ta quay nhanh cỡ nào, nhưng nhiều lỗ đen được nghiên cứu kỹ nhất dường như quay khá nhanh: vào khoảng 30% đến 100% tốc độ tối đa cho phép. Lỗ đen khổng lồ nằm ở giữa Thiên hà của chúng ta (nặng hơn Mặt Trời bốn triệu lần) có vẻ cũng là lỗ đen quay tròn, vậy nên dù chỉ 10% khối lượng của nó có thể chuyển hóa thành năng lượng hữu ích, thì năng lượng sinh ra cũng tương đương với khi 400.000 Mặt Trời được chuyển hóa thành năng lượng với hiệu suất 100%, hay tương đương với năng lượng có được từ các quả cầu Dyson bao quanh 500 triệu Mặt Trời trong hàng tỷ năm.

Quasar

Một chiến lược thú vị khác là chiết xuất năng lượng không phải từ bản thân lỗ đen, mà từ vật chất rơi vào bên trong nó. Tự nhiên đã tìm ra cách để tự mình làm vậy: đó là quasar (còn gọi là chuẩn tinh). Khi chất khí bị hút về gần lỗ đen, tạo thành một đĩa hình tròn với các

phần trong cùng dần dần bị nuốt chửng, nó trở nên cực nóng và tỏa ra lượng bức xạ vô cùng lớn. Khi chất khí rơi xuống vào trong lỗ đen, nó tăng tốc, chuyển hóa thế năng hấp dẫn thành động năng, giống như khi một người nhảy dù. Chuyển động trở nên ngày càng hỗn loạn khi nhiều động phức tạp biến chuyển động thống nhất của khối khí thành chuyển động ngẫu nhiên của các thành phần ngày một nhỏ đi, cho đến khi các nguyên tử đơn lẻ bắt đầu va chạm nhau với tốc độ cao –những chuyển động ngẫu nhiên như vậy chính là bản chất của trạng thái nóng, và những va chạm mạnh mẽ này chuyển hóa động năng thành bức xạ nhiệt. Bằng cách xây dựng một quả cầu Dyson quanh một lỗ đen, ở khoảng cách an toàn, năng lượng bức xạ này có thể được thu giữ và khai thác. Lỗ đen quay càng nhanh thì quá trình này càng hiệu quả, trong đó một lỗ đen quay ở tốc độ tối đa có thể đem lại năng lượng với hiệu suất cao đến 42%.⁵² Với những lỗ đen có khối lượng tương đương một ngôi sao, hầu hết năng lượng tỏa ra ở dạng tia X; còn với loại lỗ đen siêu khổng lồ ở trung tâm các thiên hà, hầu hết năng lượng tỏa ra ở các dạng tia hồng ngoại, ánh sáng mắt thường nhìn thấy được và tia cực tím.

Một khi đã hết nhiên liệu để nuôi lỗ đen, bạn có thể chuyển sang chiết xuất năng lượng quay tròn của nó như đã nói đến ở trên.⁵³ Quả thực, tự nhiên cũng đã tìm ra cách thực hiện phần nào điều đó, thúc đẩy bức xạ từ khối khí tích tụ qua một quá trình từ trường gọi là cơ chế Blandford- Znajek.



Hình 6.5: Theo mô hình chuẩn của vật lý hạt, 9 hạt quark với hương và spin phù hợp có thể kết hợp với nhau và biến đổi thành 3 hạt lepton qua một trạng thái trung gian gọi là sphaleron. Khối lượng kết hợp của các hạt quark (cùng với năng lượng của các hạt truyền trung gian gluon đi kèm) lớn hơn nhiều khối lượng của các lepton, nên quá trình này sẽ sinh ra năng lượng, thể hiện qua các tia phát ra.

Chúng ta hoàn toàn có thể sử dụng công nghệ để cải tiến cho hiệu suất chiết xuất năng lượng vượt quá 42% bằng cách sử dụng khéo léo từ trường hay các thành phần khác.

Sphaleron

Có một cách khác đã biết để chuyển hóa vật chất thành năng lượng mà không liên quan gì đến lỗ đen: quá trình sphaleron. Quá trình này có thể phá hủy hạt quark và biến chúng thành các hạt lepton: electron, các hạt nặng hơn như muon và tau, neutrino và các phản hạt.⁴ Như minh họa trong hình 6.5, mô hình chuẩn của vật lý hạt dự đoán rằng chín hạt quark với hương và spin phù hợp có thể kết hợp với nhau và biến đổi thành ba hạt lepton qua một trạng thái trung gian gọi là sphaleron. Vì đầu vào nặng hơn đầu ra, nên khối lượng chênh lệch

được chuyển hóa thành năng lượng theo công thức $E = mc^2$

của Einstein.

Sự sống thông minh trong tương lai do đó có thể xây dựng được cái mà tôi gọi là sphaleron hóa: một máy phát điện hoạt động như một động cơ diesel loại khủng. Một động cơ diesel truyền thống nén hỗn hợp khí và dầu diesel cho đến lúc nhiệt độ đủ cao để hỗn hợp đột nhiên bắt lửa và cháy, sau đó hỗn hợp rất nóng đó lại nở ra và trong quá trình đó làm được những việc có ích, như đẩy một cái pit-tông chẳng hạn. Khí cacbon dioxit và các khí dễ cháy khác nhẹ hơn khoảng 0,00000005% lần so với thứ ở trong pit-tông ban đầu, và chính sự chênh lệch khối lượng này biến thành nhiệt năng khiến động cơ chạy. Một sphaleron hóa lại nén vật chất bình thường nhỏ lại hàng triệu tỷ lần, rồi để nó nở ra lại và nguội đi khi sphaleron đã xong việc.⁵⁴ Ta đã biết kết quả của thí nghiệm này vì Vũ trụ sơ khai

đã thực hiện thí nghiệm cho chúng ta từ khoảng 13,8 tỷ năm trước, khi nó còn cực kỳ nóng: gần như 100% vật chất đã được chuyển hóa thành năng lượng, với chưa tới một phần tỷ của các hạt còn lại trở thành các hạt cấu tạo nên vật chất thông thường: hạt quark và hạt electron. Thế nên nó cũng như một động cơ diesel, ngoại trừ hiệu suất lớn hơn cả tỷ lần! Một lợi thế khác là bạn không cần kén chọn loại nhiên liệu – nó có hiệu quả với bất kỳ thứ gì tạo thành từ hạt quark, nghĩa là bất kỳ vật chất bình thường nào.

Nhờ những quá trình diễn ra trong nhiệt độ cao này, Vũ trụ sơ khai đã sản xuất ra lượng bức xạ (các hạt photon và neutrino) lớn hơn một nghìn tỷ lần so với vật chất (các hạt quark và electron sau đó kết hợp thành nguyên tử). Trong vòng 13,8 tỷ năm kể từ đó, một sự tách biệt lớn đã diễn ra, trong đó các nguyên tử tập trung lại thành các thiên hà, sao và hành tinh, còn hầu hết photon ở lại trong không gian giữa các thiên hà, tạo nên bức xạ phông vi sóng vũ trụ mà người ta đã dùng để tạo ra những hình ảnh về Vũ trụ sơ khai của chúng ta. Do đó, bất kỳ dạng sống bậc cao nào sống trong một thiên hà hay một khối tập trung vật chất khác đều có thể biến hầu hết vật chất mà nó đang có trở lại thành năng lượng, đưa tỷ lệ vật chất trở lại giá trị rất nhỏ ở thời Vũ trụ sơ khai bằng cách tái tạo trong thời gian ngắn những điều kiện nhiệt độ cao đậm đặc đó bên trong một Sphaleron hóa.

Để tìm hiểu hiệu suất thực tế của một sphaleron hóa, ta cần tìm ra những chi tiết thực tiễn chính: ví dụ, nó phải lớn đến đâu để ngăn một phần lớn các photon và neutrino rò rỉ ra ngoài trong giai đoạn nén? Tuy nhiên, ta có thể nói chắc chắn rằng viễn cảnh năng lượng cho sự sống tương lai tốt đẹp hơn nhiều so với những gì mà công nghệ hiện tại của chúng ta cho phép. Ta còn chưa xây được một lò phản ứng nhiệt hạch, nhưng công nghệ tương lai sẽ có thể làm được tốt hơn đến mười lần, thậm chí cả trăm lần.

Xây dựng các máy tính tốt hơn

Nếu việc ăn tối kém hiệu quả đến 10 tỷ lần so với giới hạn vật lý về hiệu suất năng lượng thì các máy tính ngày nay hiệu quả đến đâu? Như ta sẽ thấy sau đây, chúng còn kém hơn cả bữa tối đó.

Tôi thường giới thiệu Seth Lloyd, người bạn và đồng nghiệp của tôi, là người duy nhất ở MIT có thể nói là “điên” giống tôi. Sau khi

thực hiện nghiên cứu tiên phong về máy tính lượng tử, ông tiếp tục viết một cuốn sách với luận điểm rằng toàn bộ Vũ trụ là một máy tính lượng tử. Chúng tôi thường đi nhậu sau giờ làm, và tôi chưa phát hiện ra một chủ đề nào mà ông không có gì thú vị để nói. Ví dụ, như tôi đề cập trong chương 2, ông có thật nhiều điều để nói về giới hạn tận cùng của tính toán. Trong một bài viết nổi tiếng năm 2000, ông chỉ ra tốc độ tính toán bị giới hạn bởi năng lượng: thực hiện một phép tính logic cơ bản trong thời gian T cần đến năng lượng trung bình là $E = h/4T$, trong đó h là đại lượng vật lý cơ bản gọi là hằng số Planck. Tức là một máy tính nặng 1 kg có thể thực hiện tối đa 5×10^{50} phép tính trên giây – lớn gấp 1036 so với máy tính tôi đang dùng để gõ những dòng này. Ta sẽ đạt được khả năng đó trong vòng vài thế kỷ tới nếu năng lực tính toán tiếp tục nhân đôi sau mỗi hai năm, như ta đã tìm hiểu trong chương 2. Ông cũng cho thấy máy tính nặng 1 kg có thể lưu trữ tối đa 1031 bit, tức là khoảng một tỷ tỷ lần tốt hơn máy tính cá nhân của tôi.

Seth là người đầu tiên thừa nhận rằng để thực sự đạt được những giới hạn này có thể là thách thức ngay cả với sự sống siêu thông minh, vì bộ nhớ của “máy tính” 1 kg đó sẽ tương tự một vụ nổ nhiệt hạch hay một phần nhỏ của vụ nổ Big Bang. Tuy nhiên, ông lạc quan là các giới hạn thực tiễn hiện không ở xa lắm so với giới hạn tận cùng. Quả thực, các nguyên mẫu máy tính lượng tử hiện có đã thu nhỏ kích thước bộ nhớ bằng cách lưu trữ một bit trên một nguyên tử, từ đó nhân lên sẽ cho phép lưu trữ khoảng 1025 bit/kg – một ngàn tỷ lần tốt hơn máy tính của tôi. Hơn nữa, dùng bức xạ điện từ trường để kết nối giữa các nguyên tử này sẽ cho phép thực hiện khoảng 5×10^{40} phép tính trên giây – lớn gấp 1031 so với bộ xử lý trung tâm (CPU) của tôi.

Tóm lại, tiềm năng của sự sống tương lai trong việc tính toán và khám phá về mọi điều thật sự lớn tới mức khó tưởng tượng nổi: tính theo lũy thừa của 10, các siêu máy tính ngày nay còn cách chiếc máy tính 1 kg kia xa hơn cả khoảng cách tới đèn tín hiệu nhấp nháy trên một chiếc xe hơi, một thiết bị chỉ lưu trữ có một bit thông tin, thay đổi liên tục giữa bật và tắt khoảng một lần mỗi giây.

Các nguồn tài nguyên khác

Từ góc nhìn vật lý học, mọi thứ mà sự sống tương lai có thể muốn tạo ra – từ nơi ở và máy móc đến các dạng sống mới – chỉ đơn giản là các hạt cơ bản sắp xếp theo một cách đặc biệt nào đó. Cũng như cá voi xanh đơn giản là nhuyễn thể sắp xếp lại và nhuyễn thể là sinh vật phù du sắp xếp lại, toàn bộ Hệ Mặt Trời chỉ là hiđrô được sắp xếp lại trong 13,8 tỷ năm tiến hóa vũ trụ: trọng lực sắp xếp lại hiđrô thành các ngôi sao, các ngôi sao sắp xếp lại hiđrô thành các nguyên tử nặng hơn, sau đó trọng lực sắp xếp lại các nguyên tử đó thành hành tinh của chúng ta, nơi các quá trình hóa học và sinh học sắp xếp lại chúng thành sự sống.

Khi đã đạt giới hạn công nghệ, sự sống tương lai có thể thực hiện việc sắp xếp lại các hạt như thế nhanh hơn và hiệu quả hơn, bằng cách trước hết là sử dụng năng lực tính toán để tìm ra cách hiệu quả nhất, sau đó dùng năng lượng sẵn có để vận hành quá trình sắp xếp lại vật chất. Ta đã thấy cách vật chất có thể chuyển hóa thành cả máy tính và năng lượng, hay có thể hiểu vật chất là tài nguyên cơ bản duy nhất cần thiết.⁵⁵ Một khi những gì mà sự sống tương lai có thể làm với vật chất đã chạm đến các giới hạn vật lý, chỉ còn lại một cách để nó có thể làm nhiều thứ hơn: đó là kiếm thêm nhiều vật chất. Và cách duy nhất nó có thể làm thế là mở rộng ra Vũ trụ. Bay vào không gian nào!

Thu thập tài nguyên thông qua định cư vũ trụ

Tài nguyên vũ trụ của chúng ta lớn đến đâu? Cụ thể, giới hạn “trần” mà các định luật vật lý đặt ra cho lượng vật chất mà sự sống tốt cuộc có thể sử dụng được là gì? Tài nguyên vũ trụ dĩ nhiên là lớn đến khó tin, nhưng chính xác là lớn đến mức nào? Bảng 6.2 liệt kê một vài con số quan trọng. Hành tinh của chúng ta hiện giờ đã chết 99,999999%, tức là 99,999999% vật chất của Trái Đất không nằm trong sinh cầu của chúng ta và gần như không làm được gì cho đời ngoài việc cung cấp lực hấp dẫn và từ trường.

Khu vực

Số hạt

Sinh cầu của ta

1043

Hành tinh của ta

1051

<i>Hệ Mặt Trời của ta</i>	<i>1057</i>
<i>Thiên hà của ta</i>	<i>1069</i>
<i>Phạm vi di chuyển với nửa tốc độ ánh sáng</i>	<i>1075</i>
<i>Phạm vi di chuyển với tốc độ ánh sáng</i>	<i>1076</i>
<i>Vũ trụ của ta</i>	<i>1078</i>

Bảng 6.2: Số hạt vật chất (proton và neutron) xấp xỉ mà ta có thể kỳ vọng sự sống tương lai sẽ sử dụng được.

Điều này khơi gợi tiềm năng rằng một ngày nào đó chúng ta có thể sử dụng lượng vật chất gấp hàng trăm triệu lần để hỗ trợ tích cực cho sự sống. Nếu có thể khai thác tối đa tất cả vật chất trong Hệ Mặt Trời (bao gồm cả Mặt Trời), ta sẽ được kết quả tốt hơn một triệu lần nữa. Định cư trong dải Ngân Hà sẽ làm tăng nguồn tài nguyên lên một nghìn tỷ lần nữa.

Đi xa được đến đâu?

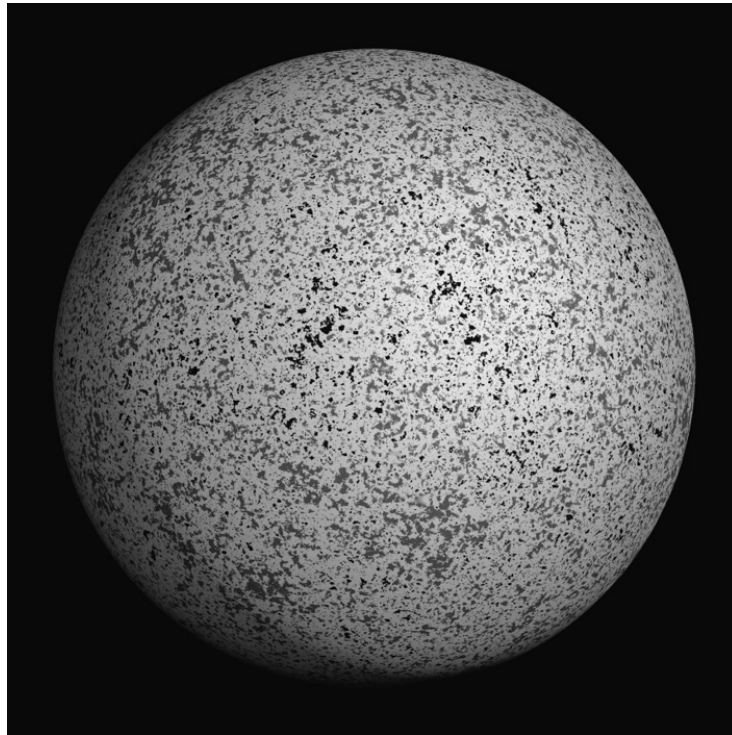
Bạn có thể cho rằng ta sẽ thu được tài nguyên vô tận bằng cách định cư ở càng nhiều thiên hà càng tốt nếu ta có đủ kiên nhẫn, nhưng vũ trụ học hiện đại không đề xuất điều đó! Phải, bản thân không gian có thể vô tận, chứa vô số thiên hà, sao và hành tinh – quả thực, đó là điều được dự đoán trong những phiên bản đơn giản nhất của

thuyết giãn nở

, mô hình khoa học phổ biến nhất hiện nay về việc cái gì đã tạo ra Vụ Nổ Lớn 13,8 tỷ năm trước. Tuy nhiên, dù số thiên hà nhiều vô hạn, nhưng các thiên hà mà ta có thể nhìn thấy và với tới thì chỉ có giới hạn: ta có thể nhìn thấy khoảng 200 tỷ và định cư ở tối đa là 10 tỷ thiên hà.

Cái giới hạn chúng ta là tốc độ ánh sáng: mỗi năm nó chỉ đi được một năm ánh sáng (khoảng 10 ngàn tỷ kilômét). Hình 6.6 cho thấy phần không gian mà từ đó ánh sáng đã đến được với chúng ta trong vòng 13,8 tỷ năm kể từ Vụ nổ lớn, một khu vực hình cầu gọi là “Vũ trụ quan sát được” hay đơn giản là “Vũ trụ của ta”. Dù không gian có là vô hạn thì Vũ trụ của ta vẫn giới hạn, “chỉ” chứa khoảng 1078 nguyên tử. Hơn nữa, khoảng 98% Vũ trụ của ta là loại “cắm sờ vào hiện vật”,

tức là ta thấy được nhưng không bao giờ chạm đến được dù có di chuyển với tốc độ ánh sáng đến muôn đời.



Hình 6.6: Vũ trụ của ta, hay khu vực hình cầu trong không gian mà từ đó ánh sáng đã có thời gian đến được với chúng ta (chúng ta ở chính giữa) trong vòng 13,8 tỷ năm kể từ Vụ Nổ Lớn. Các hoa văn là những hình ảnh của Vũ trụ sơ khai do vệ tinh Planck chụp lại, cho thấy khi mới 400.000 năm tuổi, Vũ trụ ở trạng thái vật chất plasma nóng gần bằng sức nóng trên bề mặt Mặt Trời. Không gian có lẽ còn tiếp tục trải rộng ra ngoài khu vực này, và mỗi năm ta lại thấy được thêm những vật chất mới xuất hiện trong khu vực.

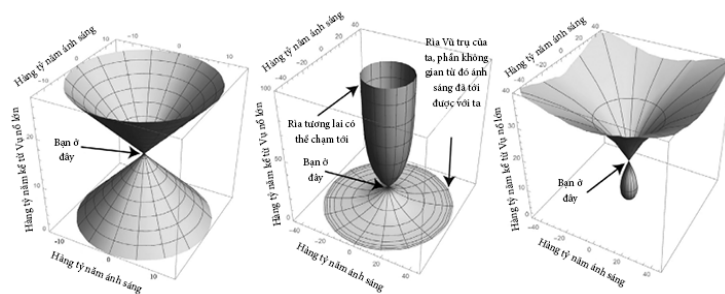
Tại sao lại như vậy? Suy cho cùng, giới hạn của những gì ta thấy được xuất phát đơn thuần từ thực tế là Vũ trụ của ta không có tuổi đời vô hạn, nên ánh sáng từ xa vẫn chưa đủ thời gian đến được với chúng ta. Vậy thì phải chăng chúng ta có thể di chuyển ngẫu hứng đến những thiên hà xa xôi nếu thời gian dành cho chuyến đi chẳng hề có giới hạn?

Thách thức đầu tiên là Vũ trụ của ta vẫn đang giãn nở, nghĩa là gần như tất cả các thiên hà đều đang bay ra xa khỏi chúng ta, nên việc định cư ở các thiên hà xa xôi trở thành trò chơi đuổi bắt. Thách

thức thứ hai là sự giãn nở vũ trụ này đang tăng tốc do ảnh hưởng của loại năng lượng tối bí ẩn cấu thành nên 70% Vũ trụ của ta. Để hiểu được tại sao điều này gây rắc rối, hãy tưởng tượng bạn đi vào ga tàu và thấy tàu đang từ từ tăng tốc đi mất, nhưng có một cửa toa vẫn đang mở ra mời gọi. Nếu nhanh chân và liều lĩnh, liệu bạn có bắt kịp con tàu không? Vì nó sẽ chạy ngày càng nhanh hơn tốc độ của bạn, nên câu trả lời rõ ràng phụ thuộc vào khoảng cách ban đầu giữa bạn với con tàu: nếu nó ở cách bạn một khoảng nhất định, bạn sẽ không bao giờ bắt kịp con tàu đó. Ta đối mặt với tình huống tương tự khi cố gắng bắt kịp các thiên hà xa xôi đang tăng tốc rời xa ta: kể cả khi có thể di chuyển với tốc độ ánh sáng thì chúng ta cũng sẽ mãi mãi không chạm tới được tất cả các thiên hà nằm bên ngoài khoảng cách 17 tỷ năm ánh sáng – và đó là hơn 98% số thiên hà trong Vũ trụ của ta rồi.

Nhưng khoan nào: chẳng phải thuyết tương đối hẹp của Einstein nói rằng không gì có thể di chuyển nhanh hơn ánh sáng hay sao? Vậy làm thế nào mà các thiên hà lại chạy nhanh hơn một thứ di chuyển với tốc độ ánh sáng được? Câu trả lời là thuyết tương đối hẹp đã được thay thế bằng thuyết tương đối rộng của chính Einstein, trong đó giới hạn tốc độ được tự do thoải mái hơn: không gì có thể di chuyển nhanh hơn tốc độ ánh sáng qua không gian, nhưng không gian thì tự do nở rộng với tốc độ tùy ý. Einstein cũng cho chúng ta một cách hay để hình dung các giới hạn tốc độ này, bằng cách coi thời gian như chiều thứ tư trong không-thời gian (xem hình 6.7, trong đó tôi đã giữ cho đồ thị ở dạng ba chiều bằng cách bỏ đi một trong các chiều không gian). Nếu không gian không nở rộng, ánh sáng sẽ tạo thành các tia ở góc 45 độ xuyên qua không-thời gian, vậy là khu vực mà ta có thể nhìn thấy và với tới từ nơi ta đang đứng có dạng hình nón. Trong khi nón ánh sáng quá khứ là nón cụt do bị cắt đứt bởi Vụ Nổ Lớn 13,8 tỷ năm trước, thì nón ánh sáng tương lai sẽ kéo dài ra mãi mãi, cho ta tiếp cận nguồn tài nguyên vũ trụ vô hạn. Trái lại, khung hình ở giữa cho thấy một vũ trụ đang nở rộng với năng lượng tối (có vẻ chính là Vũ trụ mà ta đang ở) làm biến dạng các nón ánh sáng thành hình ly rượu vang, mãi mãi giới hạn số lượng thiên hà ta có thể định cư trong khoảng 10 tỷ.

Nếu giới hạn này khiến bạn có cảm giác sợ hãi không gian hẹp của vũ trụ, hãy để tôi làm bạn vui lên với một lỗ hổng có khả năng xảy ra: tính toán của tôi giả định năng lượng tối là hằng số không thay đổi qua thời gian, thống nhất với kết luận từ những đo đạc gần đây nhất. Tuy nhiên, vì chưa thể biết chắc năng lượng tối thật sự là cái gì, nên ta vẫn còn một tia hy vọng là năng lượng tối dần dần sẽ tiêu hủy (cũng như chất tương tự năng lượng tối vốn được thừa nhận để giải thích cho sự giãn nở vũ trụ), khi ấy, sự tăng tốc sẽ nhường chỗ cho giảm tốc, có khả năng cho phép các dạng sống tương lai tiếp tục định cư ở các thiên hà mới chũng nào chúng còn tồn tại.



Hình 6.7: Trong đồ thị không-thời gian, sự kiện là một điểm mà tọa độ của nó trên các trục nằm ngang và trục thẳng đứng lần lượt xác định vị trí và thời điểm nó xảy ra. Nếu không gian không nở rộng (hình trái), thì hai hình nón tháo gỡ giới hạn các phần của không-thời gian mà chúng ta trên Trái Đất (ở đỉnh nón) có thể bị tác động (nón dưới) và gây tác động (nón trên), vì tác động nhân quả không thể đi nhanh hơn tốc độ ánh sáng, mà ánh sáng di chuyển với tốc độ một năm ánh sáng mỗi năm. Chuyện trở nên thú vị hơn khi không gian nở rộng (các hình bên phải). Theo mô hình chuẩn của vũ trụ học, ta chỉ có thể thấy và với tới một phần giới hạn của thời gian cho dù không gian là vô hạn. Trong hình giữa, trông như một ly rượu vang, ta dùng các tọa độ ẩn đi sự nở rộng của vũ trụ sao cho chuyển động của các thiên hà xa xôi qua thời gian tương ứng với các trục thẳng đứng. Từ điểm nhìn hiện tại của chúng ta, 13,8 tỷ năm sau Vụ Nổ Lớn, mới chỉ có các tia sáng từ phần đế của ly rượu có thời gian đến được với chúng ta, và dù có di chuyển với tốc độ ánh sáng thì ta cũng không bao giờ đến được những vùng bên ngoài phần trên của ly rượu, chứa khoảng 10 tỷ thiên hà. Trong hình phải, trông như giọt nước

nằm dưới một bông hoa, ta dùng các tọa độ quen thuộc cho thấy không gian nở rộng. Điều này làm biến dạng phần đế ly rượu thành hình giọt nước vì các vùng bên ngoài rìa những gì ta thấy được ban đầu nằm rất gần nhau.

Đi nhanh được đến đâu?

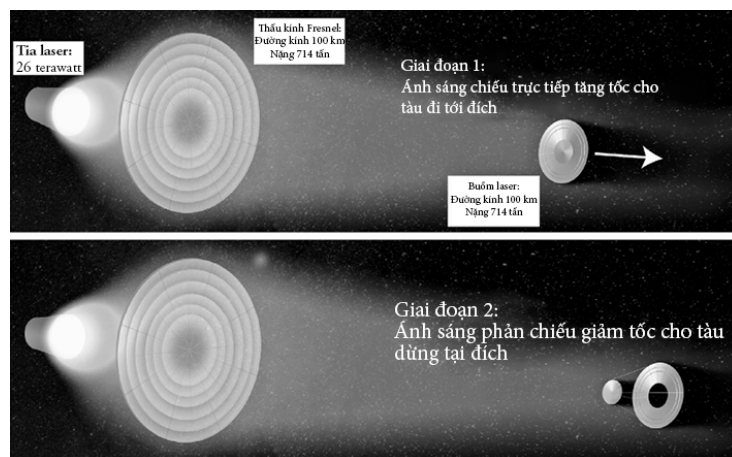
Ở trên ta đã tìm hiểu có bao nhiêu thiên hà mà một nền văn minh có thể định cư nếu nó mở rộng ra theo tất cả các hướng với tốc độ ánh sáng. Thuyết tương đối rộng nói rằng không thể bắn tên lửa qua không gian ở tốc độ ánh sáng vì như vậy sẽ cần đến năng lượng vô hạn, vậy tên lửa có thể đi nhanh đến đâu trong thực tế?⁵⁶

Tên lửa New Horizons (Chân Trời Mới) của NASA đã phá kỷ lục tốc độ khi lao tới sao Diêm Vương năm 2006 với tốc độ khoảng 100.000 dặm trên giờ (45 km trên giây), và tên lửa Solar Probe Plus năm 2018 của NASA đặt mục tiêu đi nhanh gấp hơn bốn lần bằng cách thả rơi gần Mặt Trời, nhưng ngay cả tốc độ này cũng vẫn chưa bằng 0,1% tốc độ ánh sáng. Công cuộc làm ra tên lửa ngày càng nhanh và tốt đã thu hút một số bộ óc sáng láng nhất thế kỷ qua và chúng ta có cả một kho tàng nghiên cứu phong phú và hấp dẫn về chủ đề này. Tại sao đi nhanh hơn lại khó đến thế? Có hai vấn đề chính: một là tên lửa thông thường dùng phần lớn nhiên liệu chỉ để gia tốc cho nhiên liệu mà nó mang theo; hai là nhiên liệu tên lửa ngày nay có hiệu suất thấp đến vô vọng, với tỷ lệ khối lượng chuyển thành năng lượng không lớn hơn tỷ lệ 0,00000005% của xăng mà ta thấy trong bảng 6.1 là bao. Một cách cải tiến rõ ràng là chuyển sang dùng nhiên liệu hiệu quả hơn. Ví dụ, Freeman Dyson cùng cộng sự đã nghiên cứu cho Dự án Orion của NASA, với mục tiêu cho nổ khoảng 300.000 quả bom hạt nhân trong 10 ngày để đạt được 3% tốc độ ánh sáng với một tàu vũ trụ đủ lớn để đưa loài người đến một hệ mặt trời khác trên chuyến đi dài một thế kỷ.⁵ Những người khác đã tìm hiểu việc sử dụng phản vật chất làm nhiên liệu, vì kết hợp phản vật chất với vật chất thông thường cho ra năng lượng với hiệu suất gần 100%.

Một ý tưởng phổ biến khác là làm ra tên lửa không cần mang theo nhiên liệu. Ví dụ, không gian liên sao không phải là môi trường chân không hoàn hảo, mà đôi khi chứa các ion hiđrô (một proton cô độc:

một nguyên tử hiđrô đã mất electron). Năm 1960, điều này đem lại cho nhà vật lý Robert Bussard ý tưởng xây dựng nên động cơ phản lực Bussard: phương pháp đánh bắt các ion đó trên đường đi và dùng chúng làm nhiên liệu tên lửa trong một lò phản ứng nhiệt hạch lắp trên tàu. Dù nghiên cứu gần đây đã đưa ra nghi ngờ về độ thực tiễn của phương pháp này, nhưng người ta vẫn đưa ra một ý tưởng không-mang-nhiên-liệu khác, có vẻ khả thi với một nền văn minh du hành không gian công nghệ cao: cánh buồm laser.

Hình 6.8 minh họa một thiết kế tên lửa buồm laser khéo léo được đề xướng năm 1984 bởi Robert Forward, nhà vật lý đã xây dựng nên khái niệm “vệ tinh” mà ta đã tìm hiểu cho việc xây dựng quả cầu Dyson. Cũng như các phân tử không khí va đập vào cánh buồm trên thuyền buồm sẽ đẩy thuyền đi, các hạt ánh sáng (photon) va đập vào một tấm gương sẽ đẩy nó đi. Bằng cách chiếu một tia laser khổng lồ dùng năng lượng Mặt Trời vào một cánh buồm cực lớn và cực nhẹ gắn vào tàu vũ trụ, ta có thể sử dụng năng lượng của chính Mặt Trời để gia tốc cho tên lửa đạt đến tốc độ lớn.



Hình 6.8: Thiết kế của Robert Forward cho một chuyến bay bằng buồm laser đến hệ sao α Centauri cách chúng ta bốn năm ánh sáng. Ban đầu, một tia laser cực mạnh trong Hệ Mặt Trời làm tăng tốc cho tàu vũ trụ bằng cách tạo áp lực bức xạ lên buồm laser. Để tàu chậm lại trước khi đến đích, phần rìa ngoài cánh buồm tự bung ra và phản chiếu ánh sáng laser ngược lại về phía tàu vũ trụ.

Nhưng làm cách nào để dừng lại? Đây là vấn đề mà tôi nghĩ mãi không ra cho tới khi đọc bài viết xuất sắc của Forward: như hình 6.8 cho thấy, phần rìa ngoài cánh buồm laser bung ra và di chuyển lên trước tàu vũ trụ, phản chiếu tia laser ngược lại để giảm tốc cho tàu cùng cánh buồm giờ đã thu nhỏ lại.⁶ Forward tính toán rằng làm vậy có thể cho phép con người đi một chuyến dài bốn năm ánh sáng đến hệ mặt trời α Centauri chỉ trong vòng 40 năm. Khi đã tới đó, có thể họ sẽ làm ra một hệ thống tia laser khổng lồ mới và tiếp tục nhảy từ sao này qua sao khác trên khắp Dải Ngân Hà.

Nhưng sao phải dừng lại ở đó? Năm 1964, nhà thiên văn Xô Viết Nikolai Kardashev đề xuất chấm điểm các nền văn minh theo mức năng lượng mà họ có thể khai thác để sử dụng. Khai thác được năng lượng của một hành tinh, một ngôi sao (giả sử với quả cầu Dyson) và một thiên hà lần lượt tương ứng với các nền văn minh Hạng I, Hạng II và Hạng III trên thang đo của Kardashev. Những nhà tư tưởng sau đó đã gợi ý nên có Hạng IV, tương ứng với khả năng khai thác toàn bộ Vũ trụ mà ta tiếp cận được. Từ đó, đã có cả tin vui và tin buồn cho các dạng sống tham vọng. Tin buồn là năng lượng tối tồn tại và như ta đã thấy, dường như nó giới hạn khả năng tiếp cận của chúng ta. Tin vui là sự tiến bộ phi thường của trí tuệ nhân tạo. Ngay cả những người có tầm nhìn lạc quan như Carl Sagan cũng từng coi viễn cảnh loài người đến được các thiên hà khác khá là vô vọng, khi xét đến thực tế là chúng ta sẽ chết ngay trong thế kỷ đầu tiên trên một hành trình kéo dài hàng triệu năm, kể cả nếu ta di chuyển với vận tốc gần bằng tốc độ ánh sáng. Quyết không từ bỏ, họ đã cân nhắc khả năng ướp lạnh các nhà du hành để kéo dài tuổi thọ, làm chậm quá trình lão hóa bằng cách di chuyển gần như bằng tốc độ ánh sáng, hay gửi đi cả một cộng đồng sẽ du hành qua hàng chục ngàn thế hệ nối tiếp nhau – lâu hơn cả thời gian loài người tồn tại trên Trái Đất tính tới nay.

Khả năng xuất hiện siêu trí tuệ làm thay đổi hoàn toàn bức tranh này, khiến nó trở nên hứa hẹn hơn nhiều đối với những người có máu du hành xuyên thiên hà. Tháo gỡ được nhu cầu chuyên chở các hệ thống công kênh hỗ trợ sự sống cho con người và bổ sung công nghệ do AI phát minh, định cư xuyên thiên hà bỗng nhiên có vẻ

chẳng hề phức tạp. Cánh buồm laser của Forward sẽ tốn chi phí thấp hơn khi tàu vũ trụ chỉ cần đủ lớn để chứa một “hạt giống thăm dò”: một người máy có khả năng đáp xuống một thiên thạch hay hành tinh trong hệ mặt trời mục tiêu và xây dựng một nền văn minh mới từ đầu. Nó thậm chí không cần mang theo hướng dẫn gì: tất cả việc cần làm chỉ là xây dựng một máy thu phát tín hiệu đủ lớn để nhận được những bản vẽ thiết kế và các chỉ dẫn chi tiết hơn, được truyền đi từ nền văn minh mẹ với tốc độ ánh sáng. Khi đã xong, nó sử dụng các hệ thống laser mới xây dựng để gửi đi các hạt giống thăm dò mới, mỗi lần gửi lại tiếp tục định cư ở một hệ mặt trời trong thiên hà. Ngay cả khoảng không gian dường như vô tận giữa các thiên hà cũng thường chứa một số lượng đáng kể các ngôi sao xuyên thiên hà (những ngôi sao bị đẩy ra xa khỏi thiên hà quê hương), có thể sử dụng như các trạm dừng nghỉ trên đường, hỗ trợ chiến thuật nhảy từ đảo này sang đảo khác của chuyến tàu gương buồm liên thiên hà.

Khi một hệ mặt trời hay thiên hà khác đã có AI siêu thông minh định cư, việc đưa con người đến đó thật dễ dàng – nếu con người thành công trong việc khiến AI coi đây là mục tiêu. Tất cả thông tin cần thiết về con người có thể truyền đi với tốc độ ánh sáng, sau đó AI có thể tập hợp, lắp ráp các hạt quark và electron lại, tạo ra con người mong muốn. Việc này có thể thực hiện kiểu công nghệ thấp bằng cách đơn giản là truyền đi hai gigabyte thông tin cần thiết để xác định ADN của một người, sau đó ấp trứng ra một đứa trẻ để AI nuôi lớn; hoặc AI có thể lắp ráp các hạt quark và electron bằng công nghệ nano thành người trưởng thành, có tất cả ký ức đã sao chụp từ con người nguyên bản trên Trái Đất.

Điều này nghĩa là nếu bùng nổ trí tuệ xảy ra thì vấn đề chính không phải là định cư xuyên thiên hà có khả thi không, mà chỉ là nó diễn ra nhanh đến đâu. Vì tất cả các ý tưởng ta đã tìm hiểu ở trên đều đến từ con người, chúng chỉ nên được coi là giới hạn “sàn” của tốc độ lan rộng ra vũ trụ mà sự sống có thể có; sự sống siêu thông minh đầy tham vọng có lẽ sẽ làm được tốt hơn thế nhiều và có động cơ mạnh mẽ để đẩy lùi các giới hạn, vì trong cuộc đua với thời gian và năng lượng tối, tốc độ định cư trung bình cứ tăng 1% thì sẽ có thêm 3% số thiên hà được định cư.

Ví dụ, nếu mất 20 năm để đi quãng đường dài 10 năm ánh sáng đến hệ sao tiếp theo bằng một hệ thống cánh buồm laser, rồi 10 năm nữa để ổn định và xây dựng những hệ thống laser cùng hạt giống thăm dò mới ở đó, khu vực đã định cư của không gian sẽ là một hình cầu nở ra theo mọi hướng với tốc độ trung bình bằng $1/3$ tốc độ ánh sáng. Năm 2014, trong một phân tích tuyệt vời và kỹ lưỡng về các nền văn minh tiến ra vũ trụ, nhà vật lý người Mỹ Jay Olson đã nhắc một công nghệ cao có thể thay thế cho chiến thuật nhảy đảo, bao gồm hai loại máy thăm dò riêng biệt: hạt giống thăm dò và người máy mở đường.⁷ Hạt giống thăm dò sẽ chậm lại, hạ cánh và gieo sự sống tại điểm đến. Ngược lại, người máy mở đường sẽ không bao giờ dừng chân: nó sẽ thu thập vật chất trên đường bay, có thể sử dụng một phiên bản cải tiến của công nghệ động cơ phản lực Bussard, và dùng vật chất này vừa làm nhiên liệu vừa làm nguyên liệu để chế tạo người máy mở đường và bản sao của chính nó. Đội quân mở đường tự sinh sản này sẽ liên tục tăng tốc từ tốn để luôn duy trì một tốc độ không đổi (giả sử là một nửa tốc độ ánh sáng) tương quan với các thiên hà gần đó, và sinh sản đủ thường xuyên để đội quân có thể tạo thành một lớp vỏ hình cầu nở rộng, có mật độ người máy mở đường trên diện tích vỏ luôn không đổi.

Cuối cùng là cách tiếp cận lén lút, giúp mở đường nhanh hơn bất kỳ phương pháp nào kể trên: sử dụng trò lừa đảo “rải truyền đơn vũ trụ” của Hans Moravec trong chương 4. Bằng cách phát đi một thông điệp lừa đảo, khiến những nền văn minh mới tiến hóa xây dựng nên một cỗ máy siêu thông minh để rồi cỗ máy này lại xâm chiếm chính nền văn minh non nớt đó, một nền văn minh có thể mở rộng với tốc độ gần như ánh sáng, chính là tốc độ lan tỏa trong vũ trụ của thông điệp quyến rũ mà nguy hiểm do họ phát đi. Vì đây có thể là cách duy nhất để các nền văn minh tiên tiến đi tới hầu hết các thiên hà trong nón ánh sáng tương lai của họ, và chẳng cũng không có lý do gì mà họ không nên thử cách này, nên chúng ta cần đề cao cảnh giác trước những tín hiệu gửi đến từ ngoài Trái Đất! Trong cuốn *Contact* (Liên lạc) của Carl Sagan, người Trái Đất chúng ta đã sử dụng các bản vẽ thiết kế do người ngoài hành tinh gửi đến để xây dựng một cỗ máy mà ta không hiểu rõ – tôi không khuyến khích làm vậy đâu...

Tóm lại, từ quan điểm của tôi, hầu hết các nhà khoa học và tác giả truyện khoa học viễn tưởng từng xem xét về chuyện định cư vũ trụ đều quá bi quan khi bỏ qua khả năng về siêu trí tuệ: khi giới hạn sự chú ý vào phi hành gia là con người, họ đánh giá quá mức những khó khăn của việc du hành xuyên thiên hà, còn khi giới hạn sự chú ý vào công nghệ do con người phát minh, họ lại đánh giá quá mức lượng thời gian cần thiết để đến gần giới hạn vật lý của những gì có thể xảy ra.

Giữ kết nối qua kỹ thuật vũ trụ

Nếu năng lượng tối tiếp tục gia tốc cho các thiên hà xa xôi rời xa nhau hơn như các dữ liệu thực nghiệm mới nhất cho thấy, thì tương lai của sự sống sẽ hứng chịu phiền phức đáng kể. Khi ấy, dù cho một nền văn minh tương lai có thể định cư trên hàng triệu thiên hà, nhưng năng lượng tối trải qua thời gian hàng chục tỷ năm sẽ chia nhỏ đế quốc vũ trụ này thành hàng nghìn khu vực khác nhau và không có khả năng liên lạc với nhau. Nếu sự sống tương lai không làm gì để ngăn chặn sự chia nhỏ này, thì pháo đài cuối cùng còn lại của sự sống sẽ là các quần tụ gồm khoảng một nghìn thiên hà, với tổng trọng lực đủ mạnh để đánh bại năng lượng tối đang cố chia rẽ chúng.

Điều này sẽ là động cơ lớn để thực hiện công trình kỹ thuật vũ trụ trên quy mô lớn, nếu một nền văn minh siêu thông minh muốn giữ kết nối. Nền văn minh đó sẽ có thời gian để di chuyển bao nhiêu vật chất vào bên trong quần tụ thiên hà lớn nhất trước khi năng lượng tối khiến họ mãi mãi mất kết nối với bên ngoài? Một phương pháp để di chuyển một ngôi sao qua khoảng cách lớn là thêm một ngôi sao thứ ba vào trong một hệ nhị phân gồm hai ngôi sao đang quay ổn định quanh nhau. Cũng như trong quan hệ tình cảm, khi có người thứ ba xuất hiện, tính bền vững có thể mất đi và dẫn đến một trong ba bị đẩy ra một cách thô bạo – trong trường hợp của các ngôi sao thì đẩy ra với tốc độ rất nhanh. Nếu một trong ba lại là lỗ đen, mối tình tay ba đầy biến động như thế có thể được dùng để đẩy một khối lượng đi đủ nhanh, khiến nó bay ra xa khỏi thiên hà chủ. Không may là, khi ứng dụng cho sao, lỗ đen hay thiên hà, kỹ thuật tay ba này có vẻ không có khả năng di chuyển được nhiều hơn một phần khối lượng

nhỏ của một nền văn minh đi qua quãng đường xa cần thiết để vượt mặt được năng lượng tối.

Nhưng hiển nhiên điều này không có nghĩa là sự sống siêu thông minh không thể nghĩ ra những phương pháp hay hơn, như là chuyển hóa phần lớn khối lượng ở các thiên hà xa xôi hẻo lánh thành các tàu vũ trụ có thể bay về quần tụ quê hương. Nếu có thể xây dựng sphaleron hóa, có lẽ nó còn có thể dùng để chuyển hóa vật chất thành năng lượng, chiếu vào quần tụ quê hương ở dạng ánh sáng, tại đây ánh sáng lại được tái cấu trúc thành vật chất hoặc dùng làm nguồn điện.

May mắn trên hết là nếu xây dựng được các lỗ giun bền vững có thể đi xuyên qua, cho phép liên lạc và di chuyển gần như ngay lập tức giữa hai đầu lỗ giun, bất kể chúng cách xa nhau đến đâu. Một lỗ giun là đường đi tắt trong không thời gian, giúp bạn di chuyển từ A đến B mà không đi qua không gian ở giữa. Mặc dù thuyết tương đối rộng của Einstein cho phép sự tồn tại của lỗ giun bền vững và chúng đã xuất hiện trong các bộ phim như Contact và Interstellar, nhưng chúng đòi hỏi tồn tại một loại vật chất giả thuyết kỳ lạ với khối lượng riêng âm, trong khi sự tồn tại đó có thể phụ thuộc vào những hiệu ứng hấp dẫn lượng tử mà chúng ta còn chưa hiểu rõ. Nói cách khác, các lỗ giun hữu dụng cũng có thể là bất khả thi, nhưng nếu khả thi thì sự sống siêu thông minh có động cơ to lớn để xây dựng chúng. Lỗ giun sẽ không chỉ cách mạng hóa việc liên lạc nhanh chóng bên trong các thiên hà biệt lập, mà thông qua việc kết nối các thiên hà xa xôi với quần tụ trung tâm trước đó, lỗ giun còn cho phép toàn bộ lãnh địa của sự sống tương lai duy trì kết nối trong thời gian dài, đẩy lùi hoàn toàn những nỗ lực hòng cắt đứt liên lạc của năng lượng tối. Khi hai thiên hà được kết nối bởi một lỗ giun bền vững, chúng sẽ duy trì kết nối bất kể có trôi ra xa nhau đến đâu.

Nếu, mặc cho những nỗ lực hết mình để kiến thiết vũ trụ, một nền văn minh tương lai kết luận rằng một bộ phận của mình phải chịu số phận trôi dạt ra ngoài vùng phủ sóng mãi mãi, nền văn minh ấy có thể đơn giản là để bộ phận đó trôi đi và chúc nó hạnh phúc. Tuy nhiên, nếu có những mục tiêu tính toán đầy tham vọng liên quan đến việc tìm kiếm câu trả lời cho một số câu hỏi rất khó, nó có thể chọn chiến lược đốt nung làm rẫy: chuyển hóa các thiên hà hẻo lánh thành

những máy tính khổng lồ có thể biến đổi vật chất và năng lượng thành tính toán với nhịp độ điên cuồng, với hy vọng trước khi năng lượng tối đẩy những tàn dư kiệt quệ của chúng ra khỏi tầm mắt, những máy tính này có thể truyền về cho quần tụ thiên hà mẹ những câu trả lời đã tìm kiếm từ lâu. Chiến lược đốt nung làm rẫy sẽ đặc biệt phù hợp với những khu vực xa đến nỗi chỉ có thể tới được bằng phương pháp “rải truyền đơn vũ trụ”, khiến cho những cư dân đang sinh sống ở đó phải thất vọng tràn trề. Còn ở quần tụ thiên hà quê hương, nền văn minh có thể đặt mục tiêu bảo toàn và hiệu suất tối đa để tồn tại càng lâu càng tốt.

Kéo dài được bao lâu?

Trường thọ là điều mà hầu hết những cá nhân, tổ chức và quốc gia có tham vọng đều mong muốn. Vậy nếu một nền văn minh tương lai đầy tham vọng phát triển được siêu trí tuệ và muốn trường thọ thì nó sẽ thọ được bao lâu?

Phân tích khoa học kỹ lưỡng đầu tiên về tương lai xa của chúng ta được thực hiện bởi không ai khác ngoài Freeman Dyson, và bảng 6.3 tóm tắt một số phát hiện chính của ông. Kết luận đưa ra là trừ khi trí tuệ can thiệp, các hệ mặt trời và thiên hà dần dần sẽ bị phá hủy, cuối cùng mọi thứ khác sẽ nổi gót, để lại không gì khác ngoài không gian trống rỗng, lạnh lẽo và chết chóc với một quầng sáng bức xạ mờ dần đi mãi. Nhưng Freeman kết thúc phân tích của mình bằng thái độ lạc quan: “Có những lý do khoa học tốt đẹp để nghiêm túc cân nhắc khả năng rằng sự sống và trí tuệ có thể cải tạo thành công vũ trụ của chúng ta cho mục đích riêng của mình.”⁸

Tôi nghĩ rằng siêu trí tuệ có thể dễ dàng giải quyết nhiều vấn đề liệt kê trong bảng 6.3, vì nó có thể sắp xếp lại vật chất thành thứ gì đó tốt hơn cả các hệ mặt trời và thiên hà.

<i>Sự kiện</i>	<i>Bao lâu</i>
<i>Tuổi hiện tại của Vũ trụ của ta</i>	<i>10^{10} năm</i>
<i>Năng lượng tối đẩy hầu hết thiên hà ra khỏi tầm với</i>	<i>10^{11} năm</i>
<i>Các ngôi sao cuối cùng lụi tắt</i>	<i>10^{14} năm</i>
<i>Các hành tinh tách khỏi sao</i>	<i>10^{15} năm</i>

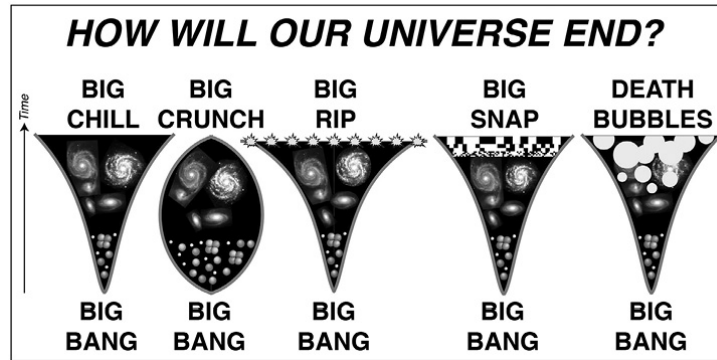
Sao tách khỏi thiên hà	10^{19} năm
Phân rã quỹ đạo do bức xạ hấp dẫn	10^{20} năm
Phân rã proton (sớm nhất)	$> 10^{34}$ năm
Các lỗ đen khối lượng tương đương sao bốc hơi	10^{67} năm
Các lỗ đen siêu khổng lồ bốc hơi	10^{91} năm
Tất cả vật chất phân rã thành sắt	10^{1500} năm
Tất cả vật chất tạo thành các lỗ đen, sau đó bốc hơi	10^{1026} năm

Bảng 6.3: Ước tính cho tương lai xa, tất cả ngoài dòng thứ 2 và thứ 7 là do Freeman Dyson tạo ra. Ông thực hiện những tính toán này trước khi người ta phát hiện ra năng lượng tối – loại năng lượng có thể gây ra nhiều dạng “tận thế vũ trụ” trong vòng 1010 tới 1011 năm tới. Proton có thể hoàn toàn bền vững; nếu không, các thí nghiệm cho thấy sẽ mất hơn 1034 năm để một nửa số proton phân rã.

Những thách thức thường được bàn đến như cái chết của Mặt Trời trong vài tỷ năm tới sẽ không phải vấn đề lớn, vì ngay cả một nền văn minh có công nghệ tương đối thấp cũng có thể dễ dàng chuyển nhà sang những ngôi sao khối lượng nhỏ có tuổi thọ hơn 200 tỷ năm. Giả sử các nền văn minh siêu thông minh xây dựng được các nhà máy điện hoạt động hiệu quả hơn các ngôi sao, hẳn họ sẽ muốn ngăn chặn sự hình thành sao để bảo toàn năng lượng: dù có dùng quả cầu Dyson để thu hoạch toàn bộ năng lượng sản sinh trong vòng đời chính của một ngôi sao (thu hồi được khoảng 0.1% tổng năng lượng), thì họ vẫn không giữ được phần lớn 99.9% năng lượng còn lại khỏi phí phạm mỗi khi những ngôi sao nặng nề chết đi. Một ngôi sao nặng chết trong một vụ nổ siêu tân tinh, hầu hết năng lượng thoát ra từ các vụ nổ này đều ở dạng các hạt neutrino khó nắm bắt; còn đối với các sao rất nặng, một phần khối lượng lớn bị phí phạm vào việc hình thành lỗ đen, mà năng lượng lại phải mất tới 1067 năm mới thoát ra được các lỗ đen này.

Chứng nào sự sống siêu thông minh còn chưa cạn kiệt nguồn vật chất/năng lượng, chứng đó nó vẫn có thể duy trì nơi ở trong trạng thái mong muốn. Có thể nó còn phát hiện ra được cách ngăn proton khỏi phân rã, sử dụng một thứ gọi là hiệu ứng Zeno trong cơ học lượng tử, trong đó quá trình phân rã bị chậm lại bằng cách liên tục quan sát quá trình. Tuy nhiên, có một nguy cơ đáng chú ý: tận thế vũ trụ có thể phá hủy toàn bộ Vũ trụ của ta chỉ trong vòng 10-100 tỷ năm nữa. Việc phát hiện ra năng lượng tối và tiến bộ trong lý thuyết dây đã làm dấy lên những viễn cảnh mới về tận thế vũ trụ mà Freeman Dyson chưa biết khi ông viết bài nghiên cứu tiên phong của mình.

Vậy Vũ trụ của ta sẽ kết thúc ra sao sau hàng tỉ năm nữa? Tôi có năm tiên liệu chính về tận thế vũ trụ sắp tới, hay “tận vũ”, như minh họa trong hình 6.9: Đại Cống, Đại Nát, Đại Xé, Đại Đứt Phứt và Bong Bóng Tử Thần. Vũ trụ của ta đã nở rộng được khoảng 14 tỷ năm. Đại Cống là khi Vũ trụ tiếp tục nở rộng mãi mãi, làm loãng không gian thành một nơi lạnh lẽo, tối tăm và cuối cùng là không còn sự sống; đây được coi là kết quả nhiều khả năng xảy ra nhất ở thời điểm Freeman viết bài nghiên cứu đó. Tôi nghĩ về nó như lời thơ T. S. Eliot: “Và như thế kết thúc cuộc đời / Bằng rên khóc chứ không bằng tiếng nổ”. Còn nếu bạn, giống như Robert Frost, mong thế giới kết thúc trong máu lửa hơn là băng giá, thì hãy cầu nguyện cho Đại Nát xảy ra, khi cuối cùng sự giãn nở vũ trụ bị đảo ngược và mọi thứ quay ra va đập vào nhau trong một vụ sụp đổ kinh hoàng gần như Vụ Nổ Lớn quay ngược. Cuối cùng, Đại Xé giống như Đại Cống cho người thiếu kiên nhẫn, khi các thiên hà, hành tinh và thậm chí các nguyên tử đều bị xé vụn trong màn kết hoành tráng sau một thời gian nhất định kể từ lúc này. Bạn nên cá cược cho viễn cảnh nào? Điều đó tùy thuộc vào việc năng lượng tối, thứ cấu thành nên 70% khối lượng Vũ trụ của ta, sẽ làm gì khi không gian tiếp tục nở rộng. Có thể là bất kỳ viễn cảnh nào trong số Cống, Nát hay Xé, tùy theo việc năng lượng tối vẫn tồn tại y nguyên không có gì thay đổi hay loãng dần thành khối lượng riêng âm, hoặc phản loãng thành khối lượng riêng lớn hơn. Vì ta vẫn chưa biết gì nhiều về năng lượng tối, nên tôi sẽ nói luôn là tôi cá thế này: 40% cho Đại Cống, 9% cho Đại Nát và 1% cho Đại Xé.



Thế còn 50% còn lại trong ví của tôi đi đâu? Tôi để dành cho lựa chọn “không cái nào đúng”, vì tôi nghĩ con người chúng ta cần khiêm tốn và công nhận là có những điều cơ bản mà ta vẫn chưa hiểu được. Chẳng hạn như bản chất của không gian. Các kết cục Cóng, Nát hay Xé đều giả định rằng không gian là bền vững và có thể kéo giãn đến vô tận. Ta từng nghĩ không gian giống như một sân khấu tĩnh tại không có gì thú vị, nơi vở kịch vũ trụ cứ thế diễn ra. Thế rồi Einstein dạy ta rằng không gian thật ra là một trong những diễn viên chính: nó có thể co lại thành lỗ đen, có thể gợn lên thành những sóng hấp dẫn và có thể kéo giãn thành một vũ trụ nở rộng. Có lẽ nó còn có thể đông cứng, như nước đông lại thành đá, thành một thể khác có những bong bóng tử thần nhanh chóng nở ra, đem tới một ứng cử viên khác có thể thay thế cho bất kỳ viễn cảnh tận thế vũ trụ nào đã kể ở trên. Nếu như bong bóng tử thần có khả năng xảy ra, có lẽ chúng sẽ nở ra với tốc độ ánh sáng, cũng như quả cầu ngày càng lớn của truyền đơn vũ trụ từ một nền văn minh hung hăng nhất.

Hơn nữa, thuyết của Einstein nói rằng giãn nở không gian có thể luôn tiếp diễn, cho phép Vũ trụ của ta tiến tới thể tích vô hạn như trong các viễn cảnh Đại Cóng và Đại Xé. Điều này nghe có vẻ quá tốt đẹp để có thể là sự thật, và tôi e là đúng như vậy. Một chiếc dây thun trông cũng tròn trịa và liên tục như không gian, nhưng nó sẽ đứt nếu bạn kéo giãn nó quá nhiều. Vì sao? Vì nó được làm từ các nguyên tử, và khi đủ giãn, bản chất của cao su là được cấu tạo từ các nguyên tử dạng hạt trở nên quan trọng. Phải chăng không gian cũng có cấu trúc dạng hạt nhưng ở quy mô đơn giản là quá nhỏ để ta chú ý? Nghiên cứu về hấp dẫn lượng tử cho thấy việc nói về không gian ba chiều truyền thống ở những quy mô nhỏ hơn 10^{-34} mét không có ý

nghĩa nhiều. Nếu đúng là không gian không thể kéo giãn vô tận mà không trải qua một vụ “Đại Đứt Phứt” khủng khiếp, thì các nền văn minh tương lai có thể sẽ muốn tái định cư trong khu vực không gian không giãn nở lớn nhất (một quần tụ thiên hà khổng lồ) mà họ có thể với tới.

Tính toán được bao nhiêu?

Sau khi khám phá sự sống tương lai

có thể

kéo dài bao lâu, hãy tìm hiểu xem liệu nó muốn kéo dài bao lâu. Con người thường muốn sống cứ càng lâu càng tốt, nhưng Freeman Dyson đưa ra một luận điểm mang tính định lượng hơn về ước muốn này: chi phí tính toán giảm đi khi bạn tính chậm, nên cuối cùng bạn sẽ làm được nhiều hơn khi bạn làm mọi thứ chậm lại hết mức có thể. Freeman còn tính được rằng nếu Vũ trụ của ta cứ mãi nở rộng và nguội dần thì một lượng tính toán vô hạn có thể là khả thi.

Chậm không hẳn là chán: nếu sự sống tương lai sống trong một thế giới mô phỏng thì trải nghiệm chủ quan về dòng chảy của thời gian không nhất thiết phải liên quan đến nhịp độ siêu chậm mà mô phỏng đó đang được chạy ở thế giới bên ngoài, thế nên viễn cảnh về tính toán vô hạn có thể tương đương với sự bất tử theo cái nhìn chủ quan của các dạng sống trong thế giới mô phỏng. Nhà vũ trụ học Frank Tipler đã dựa trên ý tưởng này để suy đoán rằng ta cũng có thể đạt được trạng thái bất tử chủ quan trong những thời khắc cuối cùng trước Đại Nát, bằng cách tăng tốc các tính toán đến giới hạn vô cực trong khi nhiệt độ và khối lượng riêng ngày càng dâng cao.

Vì năng lượng tối có vẻ phá hỏng giấc mơ tính toán vô hạn của cả Freeman và Frank, siêu trí tuệ tương lai có lẽ sẽ muốn đốt cháy các nguồn cung năng lượng của nó tương đối nhanh, để biến chúng thành tính toán trước khi gặp phải những vấn đề như chân trời vũ trụ và phân rã proton. Nếu tối đa hóa tổng lượng tính toán là mục tiêu cuối cùng thì chiến lược tốt nhất là cân bằng giữa việc quá chậm (để tránh những vấn đề nêu trên) và quá nhanh (dùng nhiều năng lượng hơn cần thiết cho mỗi tính toán).

Kết hợp mọi thứ đã tìm hiểu trong chương này, ta thấy rằng các nhà máy điện và máy tính đạt hiệu quả tối đa sẽ cho phép sự sống siêu thông minh thực hiện một lượng tính toán lớn không thể tưởng tượng nổi. Cung cấp năng lượng cho bộ não 13 watt của bạn trong 100 năm đòi hỏi năng lượng có trong khoảng nửa miligam vật chất – ít hơn cả một hạt đường bình thường. Công trình của Seth Lloyd cho thấy có thể đưa hiệu suất năng lượng của não bộ cao thêm một triệu tỷ lần, giúp hạt đường đó cung cấp được năng lượng cho mô phỏng của toàn bộ cuộc đời những con người từng sống từ xưa đến nay và hàng ngàn lần hơn thế. Nếu tất cả vật chất trong Vũ trụ hiện có của chúng ta có thể sử dụng để mô phỏng con người, chúng ta sẽ có hơn 10^{69} cuộc đời – hay bất kỳ thứ gì khác mà AI siêu thông minh muốn làm với sức mạnh tính toán của nó. Còn có thể có nhiều cuộc đời hơn nếu mô phỏng chạy chậm hơn.⁹ Ngược lại, trong cuốn Siêu trí tuệ, Nick Bostrom ước tính 10^{58} sinh mệnh con người có thể được mô phỏng với những giả định ít phóng khoáng hơn về hiệu suất năng lượng. Dù ta có phân tích những con số này như thế nào thì chúng vẫn thật là lớn, cũng lớn như trách nhiệm của chúng ta trong việc đảm bảo cho tiềm năng sinh sôi nảy nở của sự sống trong tương lai không bị phí hoài. Như Bostrom viết: “Nếu ta tượng trưng tất cả niềm hạnh phúc trải qua trong một cuộc đời như thế bằng một giọt nước mắt hạnh phúc duy nhất, thì hạnh phúc của những sinh linh này có thể liên tục đổ đầy tất cả đại dương trên Trái Đất mỗi giây, và tiếp tục như thế đến hàng trăm tỷ tỷ thiên niên kỷ nữa. Điều thật sự quan trọng là bảo đảm cho những giọt nước đó đúng là nước mắt hạnh phúc.”

Cấp bậc vũ trụ

Tốc độ ánh sáng giới hạn không chỉ sự sinh sôi mà còn cả bản chất của sự sống, đặt ra những hạn chế lớn lên truyền thông liên lạc, ý thức và kiểm soát. Vậy nếu cuối cùng phần lớn vũ trụ sẽ sống dậy thì sự sống đó sẽ như thế nào?

Cấp bậc tư duy

Bạn đã bao giờ cố đập một con ruồi bằng tay rồi thất bại chưa? Con ruồi có thể phản ứng nhanh hơn bạn là vì nó nhỏ hơn, nên cần ít thời gian hơn để thông tin truyền đi giữa mắt, não và cơ của nó. Nguyên lý “càng to = càng chậm” này áp dụng không chỉ trong sinh học, nơi giới hạn tốc độ được đặt ra bởi tốc độ truyền đi của tín hiệu xung điện qua các tế bào thần kinh, mà còn trong cả sự sống tương lai trong vũ trụ nếu không thông tin nào có thể truyền đi nhanh hơn ánh sáng. Vậy nên, đối với một hệ thống xử lý thông tin thông minh, kích thước lớn là “con dao hai lưỡi” thú vị. Một mặt, kích thước lớn giúp nó chứa được nhiều hạt hơn, cho phép tư duy phức tạp hơn. Mặt khác, điều này làm khả năng suy nghĩ thật sự tổng thể của nó chậm lại, vì thông tin liên quan cần nhiều thời gian hơn để len lỏi đến tất cả các bộ phận.

Vậy nếu sự sống bao trùm lên vũ trụ, nó sẽ chọn dạng nào: đơn giản và nhanh nhạy, hay phức tạp và chậm chạp? Tôi đoán nó sẽ lựa chọn y như sự sống trên Trái Đất đã chọn: cả hai! Các cư dân trong sinh cầu của Trái Đất có kích thước vô cùng đa dạng, từ cá voi xanh nặng 200 tấn đến vi khuẩn *Pelagibacter* siêu nhỏ chỉ 10-16 kg nhưng lại được cho là chiếm lượng sinh khối lớn hơn tất cả các loài cá trên thế giới cộng lại. Hơn nữa, các sinh vật to lớn, phức tạp và chậm chạp thường cải thiện sự “rùa bò” của mình nhờ chứa trong mình các mô-đun nhỏ hơn nhưng đơn giản và nhanh. Ví dụ, chúng ta có phản xạ chớp mắt cực kỳ nhanh chính vì nó được thực hiện bởi một đoạn mạch nhỏ và đơn giản, không liên quan đến phần lớn não bộ: nếu con ruồi khó đập đó chẳng may bay thẳng về phía mắt bạn, bạn sẽ chớp mắt trong vòng 1/10 giây, rất lâu trước khi các thông tin liên quan tới nó kịp lan truyền khắp bộ não bạn và khiến bạn nhận thức được hoàn toàn về chuyện vừa xảy ra. Nhờ tổ chức việc xử lý thông tin theo một cấp bậc các mô-đun, sinh cầu được tiện cả đôi đường, vừa có tốc độ vừa có sự phức tạp. Chính con người chúng ta cũng sử dụng chiến lược phân cấp này để tối ưu hóa việc tính toán song song.

Vì liên lạc nội bộ vừa chậm vừa tốn chi phí, nên tôi kỳ vọng rằng sự sống vũ trụ tương lai cũng sử dụng chiến thuật này, để việc tính toán được thực hiện càng cục bộ càng tốt. Nếu chỉ cần một máy tính 1 kg là đủ để thực hiện tính toán thì việc trải đều công việc này ra một

máy tính lớn cỡ thiên hà sẽ thành phần tác dụng, vì nếu cứ sau mỗi bước tính toán lại chờ đợi để thông tin được chia sẻ với tốc độ ánh sáng thì sẽ có hậu quả hết sức lố bịch là 100.000 năm trì hoãn cho mỗi bước.

Vậy liệu có phần nào trong quá trình xử lý thông tin trong tương lai này có ý thức không, tức là có bao gồm một trải nghiệm chủ quan nào hay không? Đây lại là một chủ đề thú vị gây tranh cãi mà chúng ta sẽ khám phá trong chương 8. Nếu ý thức đòi hỏi các bộ phận khác nhau trong hệ thống phải có khả năng liên lạc với nhau, thì những suy nghĩ của các hệ thống lớn hơn sẽ chậm hơn vì chúng cần phải thế. Trong khi bạn hay một siêu máy tính kích cỡ Trái Đất trong tương lai có thể có nhiều ý nghĩ trong một giây, thì một trí tuệ kích cỡ thiên hà phải mất một trăm ngàn năm mới có được một ý nghĩ, còn một trí tuệ vũ trụ có kích cỡ một tỷ năm ánh sáng thì chỉ có thời gian cho khoảng 10 ý nghĩ trước khi năng lượng tối chia nhỏ nó thành những bộ phận mất kết nối. Mặt khác, những ý nghĩ quý hiếm này và các trải nghiệm đi cùng chúng có thể khá là sâu sắc!

Cấp bậc kiểm soát

Nếu tư duy được tổ chức thành một cấp bậc có quy mô rất đa dạng, vậy còn quyền lực thì sao? Trong chương 4, ta đã khám phá cách để các thực thể thông minh tự tổ chức một cách tự nhiên thành các thứ bậc quyền lực trong cân bằng Nash, trong đó bất kỳ thực thể nào cũng sẽ khổ sở hơn nếu chúng thay đổi chiến thuật. Công nghệ truyền thông và giao thông càng tốt, các cấp bậc này càng có thể lớn hơn. Nếu siêu trí tuệ mở rộng ra quy mô vũ trụ một ngày nào đó thì cấp bậc quyền lực của nó sẽ ra sao? Sẽ là tự do và phân tán hay mang tính độc đoán cao? Sự hợp tác sẽ chủ yếu dựa trên đôi bên cùng có lợi hay ép buộc và đe dọa?

Để làm sáng tỏ các vấn đề này, hãy xem xét cả phần thưởng và hình phạt: Đây là những động cơ cho việc hợp tác trên quy mô vũ trụ, đây là những mối đe dọa có thể dùng để thi hành việc đó?

Kiểm soát bằng phần thưởng

Trên Trái Đất,

trao đổi thương mại

là động cơ truyền thống cho hợp tác vì độ khó tương đối của việc sản xuất các mặt hàng là khác nhau trên khắp hành tinh. Nếu việc khai thác 1 kg bạc ở vùng này tốn chi phí gấp 300 lần khai thác 1 kg đồng, nhưng ở vùng khác chỉ tốn gấp 100 lần, hai vùng sẽ tiến đến trao đổi 200 kg đồng với 1 kg bạc. Tương tự, nếu một vùng có công nghệ cao hơn nhiều so với vùng kia, cả hai cũng hưởng lợi từ việc trao đổi hàng hoá công nghệ cao với nguyên liệu thô.

Tuy nhiên, nếu siêu trí tuệ phát triển loại công nghệ có thể sắp sếp lại các hạt cơ bản thành bất kỳ dạng vật chất nào, điều đó sẽ loại bỏ hầu hết động cơ cho thương mại khoảng cách lớn. Sao phải vận chuyển bạc giữa các hệ mặt trời cách xa nhau trong khi chuyển đổi đồng thành bạc bằng cách sắp sếp lại cấu trúc các hạt sẽ đơn giản và nhanh hơn? Sao phải vận chuyển máy móc công nghệ cao giữa các thiên hà khi cả hiểu biết kỹ thuật và nguyên liệu thô (bất kỳ vật chất nào đều được) cùng có ở cả hai nơi? Dự đoán của tôi là trong một vũ trụ đầy nhóc siêu trí tuệ, thứ hàng hóa gần như duy nhất đáng chuyên chở qua khoảng cách lớn chính là thông tin. Ngoại lệ duy nhất có thể là vật chất để dùng cho các dự án cải tạo kỹ thuật vũ trụ – ví dụ, để chống lại nguy cơ năng lượng tối chia rẽ rồi làm phá hoại các nền văn minh như đã nói ở trên. Trái lại với thương mại truyền thống của loài người, vật chất này có thể vận chuyển ở bất kỳ dạng khối lớn nào, thậm chí có thể là một tia sáng năng lượng, vì siêu trí tuệ tiếp nhận nó có thể nhanh chóng sắp sếp lại thành bất kỳ vật thể nào nó muốn.

Nếu việc chia sẻ hay trao đổi thông tin trở thành động lực chính cho hợp tác vũ trụ thì loại thông tin nào có thể liên quan? Bất kỳ thông tin được mong muốn nào cũng sẽ có giá trị nếu việc tạo ra nó cần đến nỗ lực tính toán khổng lồ và tốn thời gian. Ví dụ, một siêu trí tuệ có thể muốn có câu trả lời cho những câu hỏi khoa học khó về bản chất của thực tại vật lý, các câu hỏi toán học khó về các định lý và thuật toán tối ưu, hay những câu hỏi kỹ thuật khó về cách tốt nhất để xây dựng công nghệ hoàn hảo. Các dạng sống thích hưởng thụ có thể muốn có những trải nghiệm mô phỏng và giải trí kỹ thuật số, và thương mại vũ trụ có thể dẫn đến nhu cầu về một loại tiền ảo vũ trụ tương tự như bitcoin.

Những cơ hội chia sẻ như vậy có thể kích thích dòng chảy thông tin không chỉ giữa các thực thể có quyền lực tương đương nhau, mà còn lên trên và xuống dưới trong bậc thang quyền lực, như giữa các trạm trung gian kích cỡ hệ mặt trời với một trung tâm thiên hà, hay giữa các trạm trung gian kích cỡ thiên hà với trung tâm vũ trụ. Các trạm trung gian có thể muốn điều này vì được là một phần của thứ gì đó lớn hơn, được cung cấp các câu trả lời và công nghệ mà họ không thể tự mình phát triển, được phòng thủ trước những mối đe dọa bên ngoài. Họ cũng đánh giá cao tiềm năng được gần như bất tử thông qua bản sao lưu: cũng như nhiều người thấy khuây khỏa nhờ tin rằng tâm trí họ sẽ sống tiếp sau khi thể xác đã chết đi, một AI cao cấp có thể trân trọng việc trí tuệ và kiến thức của nó được sống tiếp trong một siêu máy tính trung tâm sau khi phần cứng vật lý ban đầu của nó đã cạn kiệt nguồn năng lượng.

Ngược lại, trung tâm có thể muốn các trạm trung gian giúp đỡ mình trong những nhiệm vụ tính toán dài hạn khổng lồ chưa cần đến kết quả ngay, nên cũng đáng đợi hàng ngàn hay hàng triệu năm để có câu trả lời. Như ta đã tìm hiểu ở trên, trung tâm có thể muốn các trạm giúp thực hiện những dự án cải tạo kỹ thuật vũ trụ cực kỳ lớn, như chống lại sự phá hoại của năng lượng tối bằng cách di chuyển các tập hợp khối thiên hà lại với nhau. Nếu các lỗ giun đi qua được là khả thi và có thể xây dựng được, thì ưu tiên hàng đầu của trạm trung gian có lẽ sẽ là xây dựng một mạng lưới lỗ giun để ngăn năng lượng tối và giữ cho đế chế kết nối đến vô tận. Nhưng mục tiêu cuối cùng mà một siêu trí tuệ vũ trụ có thể có là gì? Đây là một vấn đề thú vị và gây tranh cãi mà ta sẽ tìm hiểu rõ hơn trong chương 7.

Kiểm soát bằng hình phạt

Các đế quốc trên Trái Đất thường buộc những kẻ dưới quyền hợp tác bằng cả phần thưởng và hình phạt. Các chư hầu của Đế quốc La Mã coi trọng công nghệ, hạ tầng và sự phòng thủ mà họ được trao cho như là phần thưởng cho việc hợp tác, đồng thời họ cũng sợ những hậu quả không thể tránh khỏi nếu họ nổi dậy hay không nộp thuế. Bởi sẽ mất nhiều thời gian để gửi các binh đoàn từ Rome ra các tỉnh xa xôi, nên một phần của sự đe dọa này được giao phó cho các binh lính địa phương và các quan chức trung thành, những người được trao quyền để đưa ra các hình phạt gần như tức

thì. Một trung tâm siêu thông minh có thể sử dụng chiến lược tương tự, duy trì một mạng lưới binh sĩ trung thành khắp đế chế vũ trụ. Vì các chư hầu siêu thông minh có thể là khó kiểm soát, nên chiến lược khả thi đơn giản nhất có lẽ là sử dụng những binh sĩ AI được lập trình với độ trung thành 100% và tương đối ngờ nghệch, chỉ làm nhiệm vụ đơn giản là theo dõi xem liệu tất cả các quy tắc có được tuân thủ hay không và nếu không thì tự động phát hỏa một thiết bị tận diệt.

Ví dụ, giả sử AI trung tâm sắp xếp cho một sao lùn trắng được đặt ở gần một nền văn minh cỡ hệ mặt trời mà nó muốn kiểm soát. Sao lùn trắng là phần còn lại của một ngôi sao tương đối nặng sau khi đã cháy hết. Cấu tạo từ phần lớn là cacbon, nó giống như một viên kim cương khổng lồ trên bầu trời, và nhỏ gọn đến nỗi nó có thể nặng hơn Mặt Trời mà vẫn nhỏ hơn Trái Đất. Nhà vật lý người Ấn Độ Subrahmanyan Chandrasekhar nổi tiếng vì đã chứng minh là nếu bạn tiếp tục tăng khối lượng cho sao lùn trắng đến khi nó vượt qua giới hạn Chandrasekhar, tức đạt khoảng 1,4 lần khối lượng Mặt Trời, nó sẽ trải qua một vụ phát nổ nhiệt hạch kinh hoàng gọi là siêu tân tinh loại 1A. Nếu AI trung tâm đã nhả tâm sắp đặt cho sao lùn trắng này cực kỳ gần với giới hạn Chandrasekhar, binh sĩ AI có thể phát huy tác dụng ngay cả khi nó rất ngờ nghệch (mà thực ra chính là vì nó rất ngờ nghệch): nó có thể được lập trình để đơn giản là xác nhận nền văn minh thuộc cấp đã nộp đủ số bitcoin vũ trụ theo quy định, các chứng minh toán học hay bất kỳ loại thuế nào đánh lên nền văn minh đó hàng tháng đã được giải quyết, và nếu không, thì quăng lên sao lùn trắng đủ khối lượng vật chất để kích nổ siêu tân tinh và làm nổ tung toàn bộ khu vực đó thành tro tàn.

Các nền văn minh kích cỡ thiên hà có thể bị kiểm soát tương tự bằng cách đặt nhiều vật thể nhỏ gọn vào các quỹ đạo chặt chẽ quanh lỗ đen khổng lồ ở tâm thiên hà, rồi đe dọa biến đổi các khối này thành khí, ví dụ bằng cách cho chúng đâm vào nhau. Khí này sẽ làm đầy lỗ đen, biến nó thành một quasar cực mạnh, có khả năng khiến cả thiên hà không còn là nơi sinh sống được nữa.

Tóm lại, có những động cơ mạnh mẽ để sự sống tương lai hợp tác qua những khoảng cách lớn trong vũ trụ, nhưng liệu sự hợp tác đó sẽ chủ yếu dựa trên lợi ích đôi bên hay đe dọa bạo lực thì vẫn là

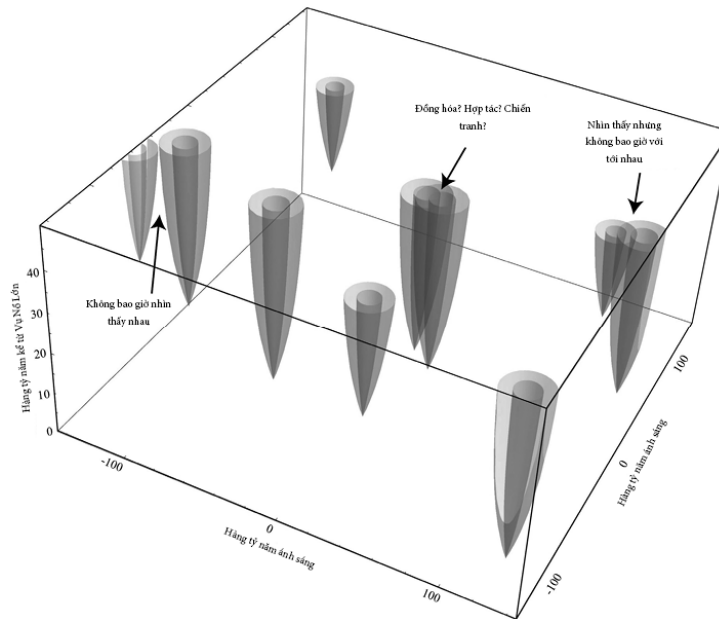
một câu hỏi mở – các giới hạn mà vật lý đặt ra dường như cho phép cả hai tình huống, nên kết quả sẽ phụ thuộc vào các mục tiêu và giá trị chính yếu. Ta sẽ tìm hiểu khả năng chúng ta có thể tác động lên các mục tiêu và giá trị của sự sống tương lai trong chương 7.

Khi các nền văn minh va chạm

Cho tới lúc này, chúng ta chỉ bàn đến các viễn cảnh sự sống mở rộng ra vũ trụ từ một bùng nổ trí tuệ duy nhất. Nhưng điều gì sẽ xảy ra khi sự sống tiến hóa độc lập ở nhiều nơi và hai nền văn minh đang mở rộng gặp nhau?

Hệ mặt trời bất kỳ nào cũng có xác suất là sự sống sẽ tiến hóa trên một trong các hành tinh của nó, phát triển công nghệ tiên tiến và mở rộng ra vũ trụ. Xác suất này có vẻ lớn hơn 0 vì sự sống công nghệ đã tiến hóa tại đây, trong Hệ Mặt Trời của chúng ta, và các định luật vật lý dường như cho phép diễn ra việc định cư trong không gian. Nếu không gian đủ lớn (trong thực tế, lý thuyết giãn nở vũ trụ cho rằng không gian là cực kỳ lớn hay vô tận), thì sẽ có nhiều nền văn minh mở rộng như thế, như minh họa trong hình 6.10. Bài nghiên cứu đề cập ở trên của Jay Olson bao gồm một phân tích rất hay về những sinh cầu vũ trụ dần mở rộng, và Toby Ord cũng đã thực hiện phân tích tương tự với các cộng sự tại Viện Tương lai Nhân loại. Xét trong không gian ba chiều, các sinh cầu vũ trụ này là những hình cầu theo đúng nghĩa đen, miễn là các nền văn minh mở rộng với cùng tốc độ về mọi hướng. Trong không thời gian, trông chúng giống như phần trên của ly rượu vang trong hình 6.7, vì cuối cùng năng lượng tối sẽ giới hạn số thiên hà mỗi nền văn minh có thể chạm tới.

Nếu khoảng cách giữa các nền văn minh láng giềng định cư trong không gian lớn hơn nhiều so với mức độ mở rộng mà năng lượng tối cho phép, thì các nền văn minh đó sẽ không bao giờ chạm mặt nhau hay thậm chí biết được về sự tồn tại của nhau, nên họ sẽ thấy như mình đơn độc trong vũ trụ.



Hình 6.10: Nếu sự sống tiến hóa độc lập ở nhiều điểm trong không thời gian (địa điểm và thời điểm) và bắt đầu xâm lấn ra không gian, thì không gian sẽ chứa một mạng lưới gồm các sinh cầu vũ trụ đang mở rộng, mỗi sinh cầu trông giống như phần trên của ly rượu vang trong hình 6.7. Đáy của mỗi sinh cầu đại diện cho địa điểm và thời điểm mà sự xâm lấn bắt đầu. Các ly rượu đục và trong lần lượt tượng trưng cho tốc độ xâm lấn bằng 50% và 100% tốc độ ánh sáng, còn các ly rượu giao nhau cho thấy những nơi các nền văn minh độc lập gặp nhau.

Tuy nhiên, nếu vũ trụ của ta chật hẹp hơn khiến cho các hàng xóm láng giềng gần nhau hơn thì cuối cùng, một số nền văn minh sẽ giao nhau. Chuyện gì xảy ra trong các khu vực giao nhau này? Sẽ là hợp tác, cạnh tranh hay chiến tranh?

Người châu Âu từng chinh phục được châu Phi và châu Mỹ vì họ có công nghệ cao hơn. Ngược lại, có thể là từ lâu trước khi hai nền văn minh siêu thông minh gặp nhau, công nghệ của họ đã phát triển ngang bằng nhau, chỉ bị giới hạn bởi các định luật vật lý. Điều này khiến ít có khả năng một siêu trí tuệ dễ dàng chinh phục siêu trí tuệ khác dù có muốn. Hơn nữa, nếu các mục tiêu của họ đã tiến hóa trở nên tương đối hài hòa thì có thể họ sẽ ít có mong muốn chinh phục hay chiến tranh. Ví dụ, nếu họ đều đang cố gắng chứng minh càng nhiều định lý tuyệt vời càng tốt và phát minh các thuật toán thông

minh nhất có thể, họ chỉ cần chia sẻ những tìm hiểu của mình và cả hai sẽ cùng phát triển. Suy cho cùng, thông tin rất khác với những tài nguyên mà con người thường tranh chấp, vì nó có thể đồng thời vừa cho đi vừa giữ lại.

Một số nền văn minh mở rộng có thể có những mục tiêu về bản chất là không thay đổi được, như các mục tiêu của một giáo phái chính thống hay một virus lây lan. Tuy nhiên, cũng có khả năng một số nền văn minh cao cấp sẽ giống như những con người suy nghĩ rộng mở hơn – sẵn sàng điều chỉnh mục tiêu khi có các lập luận đủ sức thuyết phục được đưa ra. Nếu hai nền văn minh đó gặp nhau, sẽ có một sự đụng độ không phải bằng vũ khí mà bằng tư tưởng, trong đó bên nào thuyết phục hơn sẽ thắng, và các mục tiêu của bên đó được lan truyền với tốc độ ánh sáng qua khu vực được bên còn lại kiểm soát. Đồng hóa láng giềng là một chiến lược mở rộng nhanh hơn cả định cư, vì tầm ảnh hưởng có thể lan ra với tốc độ di chuyển của ý tưởng (tốc độ ánh sáng sử dụng công nghệ viễn thông), trong khi định cư vật lý không thể tránh khỏi việc diễn ra chậm hơn tốc độ ánh sáng. Sự đồng hóa không mang tính ép buộc như cách làm tai tiếng của nhân vật Borg trong bộ phim khoa học viễn tưởng Star Trek, mà là tự nguyện, dựa trên sức thuyết phục mạnh mẽ của tư tưởng, giúp bên bị đồng hóa được tác động tích cực.

Ta đã thấy là vũ trụ tương lai có thể chứa hai loại bong bóng nhanh chóng mở rộng: các nền văn minh tiến ra vũ trụ và các bong bóng tử thần nở ra với tốc độ ánh sáng và phá hủy tất cả các hạt cơ bản, khiến không gian trở thành nơi không sinh sống được nữa. Một nền văn minh tham vọng do đó có thể gặp phải ba loại khu vực khác nhau: khu vực không ai sinh sống, bong bóng sự sống và bong bóng tử thần. Nếu lo sợ những nền văn minh đối địch không hợp tác, nó sẽ có động cơ để bắt đầu công cuộc “chiếm đất” nhanh chóng và định cư ở các khu vực hoang vắng trước đối thủ. Tuy nhiên, ngay cả khi không có nền văn minh nào khác, nó cũng có động cơ mở rộng tương tự, đơn giản là để lấy được các tài nguyên trước khi năng lượng tối khiến chúng không chạm tới được nữa. Ta vừa thấy việc chạm mặt một nền văn minh khác có thể hoặc tốt hơn hoặc tệ hơn việc gặp được không gian hoang vu, tùy xem hàng xóm hợp tác và tư duy mở đến đâu. Tuy nhiên, gặp mặt bất kỳ nền văn minh đang

xâm lấn nào (kể cả một nền văn minh cố gắng biến cư dân của bạn thành kếp giấy) vẫn còn hơn là gặp phải bong bóng tử thần, thứ sẽ tiếp tục nở ra với tốc độ ánh sáng bất kể bạn có cố gắng chống lại hay thuyết phục nó thế nào. Thứ duy nhất bảo vệ chúng ta trước bong bóng tử thần là năng lượng tối: nó ngăn không cho những bong bóng xa xôi chạm tới chúng ta. Nên nếu bong bóng tử thần quả thực là không hiếm gặp, thì năng lượng tối thực ra không phải là kẻ thù mà là bạn ta.

Ta có một mình không?

Nhiều người mặc nhiên công nhận là có sự sống cao cấp trên khắp Vũ trụ của ta, nên nếu loài người tuyệt chủng thì đó cũng không phải vấn đề đáng kể khi xét trên góc độ toàn vũ trụ. Suy cho cùng, sao ta phải lo lắng về chuyện xóa sổ nhân loại nếu một nền văn minh thú vị nào đó như

Star Trek

sẽ mau chóng tiến vào và gieo lại hạt mầm sự sống trong Hệ Mặt Trời, thậm chí có thể sử dụng công nghệ tiên tiến của họ để tái tạo và hồi sinh chúng ta? Tôi coi giả định

Star Trek

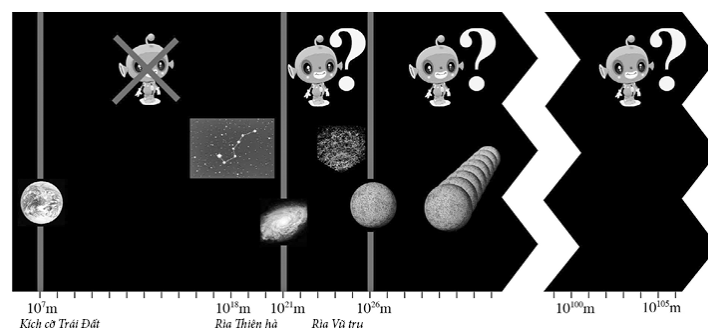
này là nguy hiểm, vì nó có thể đẩy chúng ta vào cảm giác an toàn giả mạo, khiến cho nền văn minh của ta thờ ơ hay mất thận trọng trước viễn cảnh tương lai. Hơn nữa, tôi nghĩ giả định rằng ta không chỉ có một mình trong vũ trụ không những nguy hiểm mà còn có thể là sai.

Quan điểm này của tôi là thiếu sót⁵⁷ và hoàn toàn có thể sai lầm, nhưng ít nhất đó là một khả năng mà hiện tại ta không thể bỏ qua. Nó đem lại cho chúng ta lời khẩn cầu về mặt đạo đức là phải giữ an toàn và không để cho văn minh nhân loại diệt vong.

Khi giảng về vũ trụ học, tôi thường yêu cầu người nghe giơ tay nếu họ nghĩ có sự sống thông minh nào khác trong Vũ trụ của ta (khu vực trong không gian mà ánh sáng từ đó đã tới được với ta trong vòng 13,8 tỷ năm kể từ Vụ Nổ Lớn). Lần nào cũng như lần nào, gần như tất cả đều giơ tay, từ trẻ mẫu giáo tới sinh viên đại học. Khi

tôi hỏi vì sao, câu trả lời cơ bản tôi thường nhận được là Vũ trụ của ta rộng lớn như thế thì hẳn phải có sự sống ở đâu đó, ít nhất là về mặt thống kê. Hãy xem xét kỹ hơn lập luận này và chỉ ra điểm yếu của nó.

Tất cả đều nằm ở một con số: khoảng cách điển hình giữa một nền văn minh trong hình 6.10 với láng giềng gần nhất. Nếu khoảng cách này lớn hơn nhiều so với 20 tỷ năm ánh sáng, ta nên kỳ vọng là chỉ có một mình trong Vũ trụ của ta (khu vực trong không gian mà ánh sáng từ đó đã tới được với ta trong vòng 13,8 tỷ năm kể từ Vụ Nổ Lớn), và không bao giờ liên lạc được với người ngoài hành tinh. Vậy ta nên ước chừng khoảng cách này là bao nhiêu? Chúng ta gần như không có manh mối nào. Nghĩa là khoảng cách tới hàng xóm của chúng ta nằm trong khoảng 1000 ... 000 mét, mà số các con số 0 có thể là bất kỳ số nào từ 21, 22, 23, ..., 100, 101, 102 hay hơn thế nữa – nhưng có lẽ không nhỏ hơn 21 là bao, vì ta vẫn chưa thấy bằng chứng thuyết phục nào về người ngoài hành tinh (xem hình 6.11). Để nền văn minh láng giềng gần nhất có bán kính khoảng 10^{26} mét nằm trong Vũ trụ của ta, số các số 0 không thể vượt quá 26, mà xác suất để số các số 0 rơi vào khoảng hẹp giữa 22 và 26 lại khá là nhỏ. Đó là lý do tôi nghĩ ta chỉ có một mình trong Vũ trụ của ta.



Hình 6.11: Ta có một mình không? Những phỏng đoán còn thiếu nhiều điều chắc chắn về cách sự sống và trí tuệ tiến hóa cho thấy nền văn minh láng giềng gần nhất của chúng ta có thể nằm ở bất kỳ đâu trên trục ngang phía trên, khiến nó khó có khả năng nằm ở khoảng không hẹp giữa rìa Thiên hà của ta (cách ta khoảng 1021 mét) và rìa Vũ trụ của ta (cách ta khoảng 1026 mét). Nếu nó ở gần hơn khoảng này nhiều, thì tức là phải có nhiều nền văn minh tiên tiến

khác trong Thiên hà mà chúng ta hẳn đã thấy được, nên điều này cho thấy thực ra ta chỉ có một mình trong Vũ trụ của ta.

Tôi đã đưa ra các chứng minh chi tiết về luận điểm này trong cuốn Vũ trụ toán học của chúng ta, nên tôi sẽ không nhắc lại ở đây, nhưng lý do cơ bản khiến ta không có ý niệm gì về khoảng cách với láng giềng là ta cũng không biết gì hết về xác suất có sự sống thông minh nảy sinh ở một nơi nhất định. Như nhà thiên văn học người Mỹ Frank Drake đã chỉ ra, xác suất này có thể tính được bằng cách nhân xác suất có môi trường sinh sống được (chẳng hạn một hành tinh phù hợp) với xác suất sự sống sẽ hình thành ở đó và xác suất sự sống đó sẽ tiến hóa trở thành thông minh. Khi tôi còn là sinh viên đại học, chúng ta chưa biết được gì về ba xác suất này. Sau những phát hiện kỳ tích trong hai thập kỷ qua về các hành tinh quay quanh các ngôi sao khác, khả năng hiện nay là có rất nhiều hành tinh sống được và ngay trong chính Thiên hà của chúng ta cũng có hàng tỷ hành tinh như thế. Tuy nhiên, xác suất hình thành sự sống và sự sống tiến hóa thành thông minh thì vẫn còn cực kỳ thiếu chắc chắn: một số chuyên gia cho rằng một hay cả hai khả năng này là khó tránh khỏi và xảy ra ở hầu hết các hành tinh sinh sống được, trong khi người khác lại nghĩ một hay cả hai khả năng đều cực kỳ hiếm gặp vì phải may mắn vô cùng thì chúng mới lọt qua được một hay nhiều nút thắt cổ chai của quá trình tiến hóa. Người ta đề xuất một số nút thắt cổ chai như vấn đề gà-hay-trứng-có-trước ở những giai đoạn đầu của sự sống tự sinh sản: ví dụ, để một tế bào hiện đại xây dựng một ribosome – cỗ máy phân tử hết sức phức tạp có khả năng đọc mã gien di truyền và xây dựng các protein – nó cần một ribosome khác, và chúng ta còn chưa rõ rằng ribosome đầu tiên có khả năng tiến hóa dần dần từ một thứ đơn giản hơn hay không.¹⁰ Một số nút thắt cổ chai khác lại liên quan tới sự phát triển của trí tuệ bậc cao. Ví dụ, mặc dù khủng long thống trị Trái Đất trong hơn 100 triệu năm, khoảng thời gian dài gấp 1000 lần so với sự tồn tại của loài người tính đến nay, nhưng dường như tiến hóa không hề đẩy chúng tiến tới trí tuệ bậc cao để chúng phát minh ra kính viễn vọng hay máy tính.

Một số người phản bác lập luận của tôi bằng cách nói rằng, phải, sự sống thông minh có thể rất hiếm, nhưng thực ra là không – Thiên

hà của chúng ta chứa đầy sự sống thông minh mà các nhà khoa học chính thống không nhận ra. Có thể người ngoài hành tinh đã đến thăm Trái Đất, như những người tin vào UFO khẳng định. Có thể người ngoài hành tinh chưa đến Trái Đất, nhưng họ ở ngoài đó và chủ động trốn tránh chúng ta (đây được gọi là “giả thuyết vườn thú”, do nhà thiên văn học người Mỹ John A. Ball đặt tên và xuất hiện trong những tác phẩm khoa học viễn tưởng kinh điển như Star Maker của Olaf Stapledon). Hoặc có thể họ ở ngoài đó mà không cố ý trốn tránh: họ đơn giản là không quan tâm tới việc định cư ngoài vũ trụ hay những dự án kỹ thuật lớn mà ta đã nhận thấy.

Hẳn rồi, ta cần có tư duy rộng mở về những khả năng này, nhưng vì không có bằng chứng nào được công nhận rộng rãi nên ta cũng cần nghiêm túc cân nhắc khả năng còn lại rằng ta chỉ có một mình. Hơn nữa, tôi nghĩ ta không nên đánh giá thấp sự đa dạng của các nền văn minh bậc cao bằng cách giả định rằng họ đều có chung những mục tiêu khiến họ không bị chú ý: ta đã thấy ở trên rằng việc thu thập tài nguyên là một mục tiêu khá hiển nhiên mà một nền văn minh cần có, và để khiến ta chú ý thì chỉ cần một nền văn minh quyết định tiến ra định cư ở mọi nơi có thể, càn quét Thiên hà của ta và xa hơn nữa. Nếu quả thực có hàng triệu hành tinh sinh sống được như Trái Đất trong Thiên hà và có tuổi đời lâu hơn Trái Đất hàng tỷ năm, đem lại dư thời gian cho các cư dân tham vọng định cư trong Thiên hà, thì hẳn ta phải nhận biết được. Do đó, chúng ta càng không thể bỏ qua cách giải thích rõ ràng nhất: rằng khởi nguồn sự sống đòi hỏi một may mắn ngẫu nhiên hiếm có tới mức tất cả các hành tinh đó đều không có sự sống.

Nếu sự sống không hiếm gặp, thì sớm muộn ta cũng sẽ biết. Các khảo sát thiên văn đầy tham vọng đang tìm kiếm trong khí quyển của các hành tinh như Trái Đất bằng chứng về khí oxy do sự sống sinh ra. Song song với việc tìm kiếm bất kỳ dấu hiệu nào của sự sống, cuộc tìm kiếm sự sống thông minh gần đây được đẩy mạnh bằng dự án trị giá 100 triệu đô-la có tên “Đột phá Lắng nghe” của nhà hảo tâm người Nga Yuri Milner.

Điều quan trọng là chúng ta không quá đề cao loài người khi tìm kiếm sự sống bậc cao: nếu ta phát hiện ra một nền văn minh ngoài Trái Đất, nhiều khả năng nó đã có siêu trí tuệ rồi. Như Martin Rees

viết trong một bài luận gần đây: “lịch sử của nền văn minh công nghệ nhân loại được tính bằng thế kỷ – và có thể chỉ một hay hai thế kỷ nữa, trước khi loài người bị thống trị hay vượt qua bởi trí tuệ phi hữu cơ, và rồi trí tuệ đó sẽ duy trì, tiếp tục tiến hóa, thêm hàng tỷ năm nữa ... Ta sẽ ít khả năng “bắt ” được nó trong khoảng thời gian ngắn ngủi mà nó tồn tại ở dạng hữu cơ.”¹¹ Tôi đồng ý với kết luận của Jay Olson trong bài nghiên cứu về định cư không gian đã nhắc đến ở trên: “Chúng tôi coi khả năng sự sống bậc cao sẽ sử dụng tài nguyên của vũ trụ để đơn giản là đưa các phiên bản tiên tiến của loài người lấp đầy các hành tinh như Trái Đất và đang tồn tại là một đích đến khó xảy ra của tiến bộ công nghệ.” Nên khi bạn tưởng tượng ra người ngoài hành tinh, đừng nghĩ đến những anh bạn màu xanh lá cây nhỏ bé với hai chân hai tay, mà hãy nghĩ về sự sống du hành vũ trụ siêu thông minh mà ta đã tìm hiểu trong chương này.

Dù là người ủng hộ mạnh mẽ tất cả các cuộc tìm kiếm sự sống ngoài Trái Đất đang cố gắng làm sáng tỏ một trong những câu hỏi thú vị nhất của khoa học, nhưng tôi vẫn ngẫm hy vọng rằng tất cả sẽ thất bại và không tìm ra gì hết! Sự khập khiễng hiển hiện giữa việc có rất nhiều hành tinh sinh sống được trong Thiên hà với việc không có vị khách ngoài Trái Đất nào, gọi là nghịch lý Fermi, cho thấy sự tồn tại của cái mà nhà kinh tế học Robin Hanson gọi là “Bộ lọc lớn”, một chướng ngại vật về mặt tiến hóa/công nghệ ở đâu đó trên con đường phát triển từ vật chất không sự sống thành sự sống định cư trong vũ trụ. Nếu ta phát hiện sự sống tiến hóa độc lập ở đâu đó khác, thì điều này có nghĩa là sự sống khởi thủy không hiếm và chướng ngại vật nằm phía sau giai đoạn phát triển hiện tại của loài người – có lẽ vì định cư vũ trụ là bất khả thi, hoặc vì hầu hết các nền văn minh bậc cao đều tự hủy diệt trước khi có thể tiến ra vũ trụ. Đó là lý do vì sao tôi cầu cho tất cả các cuộc tìm kiếm sự sống ngoài Trái Đất không tìm ra gì hết: điều này nhất quán với tình huống là sự sống thông minh tiến hóa rất hiếm nhưng con người chúng ta may mắn nên chướng ngại vật nằm phía sau chúng ta và ta vẫn có tiềm năng tương lai vô cùng to lớn.

Triển vọng

Chúng ta đã dành từ đầu cuốn sách tới giờ cho việc tìm hiểu lịch sử sự sống trong Vũ trụ của ta, từ thuở sơ khai hàng tỷ năm trước đến những khả năng trong tương lai hàng tỷ năm sau. Nếu sự phát triển AI hiện tại cuối cùng sẽ dẫn đến bùng nổ trí tuệ và tối ưu hóa định cư vũ trụ, sẽ có một cuộc bùng nổ thực sự trên toàn vũ trụ: sau hàng tỉ năm tồn tại như một nhiễu loạn nhỏ bé không đáng kể giữa một vũ trụ vô tri vô giác, sự sống bỗng nhiên bùng phát ra đầy trường vũ trụ như một luồng sóng hình cầu lan ra gần như với tốc độ ánh sáng, không bao giờ chậm lại, và khiến mọi thứ trên đường đi của nó bùng lên ngọn lửa sự sống.

Quan điểm lạc quan như vậy về tầm quan trọng của sự sống đối với tương lai vũ trụ đã được trình bày hùng hồn bởi nhiều nhà tư tưởng đã được nhắc đến trong cuốn sách. Vì các tác giả truyện khoa học viễn tưởng thường bị coi như những kẻ mơ mộng lãng mạn thiếu thực tế, tôi thấy thật trớ trêu là hầu hết các áng văn khoa học và khoa học viễn tưởng viết về định cư vũ trụ giờ đây đều có vẻ quá bi quan dưới ánh sáng của siêu trí tuệ. Ví dụ, ta đã thấy việc du hành xuyên thiên hà trở nên dễ dàng hơn nhiều một khi con người và các thực thể thông minh khác có thể được truyền đi dưới dạng số, khiến chúng ta có tiềm năng trở thành chủ nhân số phận của mình không chỉ trong Hệ Mặt Trời hay Dải Ngân Hà, mà trên toàn vũ trụ.

Ở trên ta đã cân nhắc khả năng hết sức thực tế rằng chúng ta là nền văn minh công nghệ cao duy nhất trong Vũ trụ của ta. Hãy dành phần còn lại của chương này khám phá tình huống này, và trách nhiệm đạo đức to lớn mà nó đặt ra. Điều này nghĩa là sau 13,8 tỷ năm, sự sống trong Vũ trụ của ta đã đi đến ngã ba đường, đối mặt với lựa chọn giữa sinh sôi trên khắp vũ trụ hay tuyệt chủng. Nếu ta không tiếp tục cải tiến công nghệ, vấn đề không còn là liệu nhân loại có diệt vong không, mà là diệt vong bằng cách nào. Điều gì sẽ xảy đến với chúng ta trước – một thiên thạch, siêu núi lửa phun trào, sức nóng hủy diệt của Mặt Trời khi già đi, hay một tai ương nào khác (xem hình 5.1)? Một khi ta không còn trên đời, màn kịch vũ trụ theo như Freeman Dyson dự đoán sẽ tiếp diễn mà không có khán giả: khi đối diện với tận thế vũ trụ, các ngôi sao sẽ cháy hết, các thiên hà tắt dần và các lỗ đen bốc hơi, tự kết liễu đời mình bằng một vụ nổ cực lớn tỏa ra năng lượng gấp một triệu lần năng lượng của Tsar Bomba,

quả bom hiđrô mạnh nhất từng được chế tạo. Như Freeman viết: “Vũ trụ lạnh lẽo đang nở rộng sẽ được chiếu sáng bằng những trận pháo hoa thăng hoa hoặc trong một thời gian rất dài.” Trời ơi, màn bắn pháo hoa này sẽ là sự phí phạm vô nghĩa, khi chẳng còn ai trên đời để thưởng thức.

Không có công nghệ, sự tuyệt chủng của loài người sẽ tới rất sớm trong bối cảnh vũ trụ của hàng chục tỷ năm nữa, khiến màn kịch sự sống trong Vũ trụ của ta chỉ còn là một tia sáng ngắn ngủi thoáng qua của cái đẹp, lòng đam mê và ý nghĩa giữa sự vô nghĩa gần như vĩnh cửu và không có ai trải nghiệm. Quả là sự phí phạm một cơ hội trời cho! Nếu như ta chọn cách đón nhận công nghệ thay vì né tránh nó, thì tức là ta đánh cược lớn hơn: ta sẽ được hai khả năng, một là sự sống sinh tồn và nảy nở, hai là sự sống diệt vong sớm hơn, tự hủy diệt do lên kế hoạch kém (xem hình 5.1). Tôi bỏ phiếu cho việc đón nhận công nghệ, và tiến hành những thứ mà ta chế tạo, không phải bằng niềm tin mù quáng, mà bằng sự thận trọng, tiên đoán và lập kế hoạch cẩn thận.

Sau 13,8 tỷ năm lịch sử vũ trụ, ta thấy mình đang ở trong một Vũ trụ đẹp đến ngạt thở, và thông qua loài người chúng ta, vũ trụ ấy đã sống dậy và bắt đầu tự nhận thức về mình. Ta đã thấy tiềm năng tương lai cho sự sống trong Vũ trụ của ta to lớn hơn cả những giấc mơ hoang đường nhất của tổ tiên ta, đối trọng bằng một khả năng cũng thực tế không kém là sự sống thông minh diệt vong mãi mãi. Liệu sự sống trong Vũ trụ của ta sẽ hoàn thiện hay bỏ phí tiềm năng của nó? Điều này dựa một phần lớn vào những gì con người chúng ta đang sống hôm nay làm được trong cuộc đời, và tôi lạc quan rằng chúng ta có thể khiến cho tương lai của sự sống thực sự tuyệt vời nếu ta lựa chọn đúng. Ta nên mong muốn điều gì và làm thế nào để đạt được những mục tiêu đó? Hãy dành phần còn lại của cuốn sách để tìm hiểu một số thách thức khó khăn nhất có liên quan và ta có thể làm gì với những thách thức đó.

TÓM TẮT

So với thang thời gian hàng tỷ năm của vũ trụ, bùng nổ trí tuệ là một sự kiện đột ngột khi công nghệ nhanh chóng đạt trạng thái bình ổn ở một cấp độ chỉ bị giới hạn bởi các định luật vật lý.

Trạng thái bình ổn công nghệ này ở cấp độ cao hơn nhiều công nghệ ngày nay, cho phép một lượng vật chất nhất định sinh ra năng lượng lớn gấp hàng chục tỷ lần (sử dụng sphaleron hay lỗ đen), chứa được lượng thông tin lớn gấp 1012 đến 1018 lần, hay tính toán nhanh hơn gấp 1031 đến 1041 lần – hoặc được chuyển hóa thành bất kỳ dạng vật chất mong muốn nào.

Sự sống siêu thông minh không chỉ sử dụng hiệu quả hơn đáng kể các nguồn tài nguyên sẵn có, mà còn có khả năng phát triển sinh cầu hiện nay lớn thêm 1032 lần bằng cách thu thập thêm tài nguyên từ định cư vũ trụ gần như với tốc độ ánh sáng.

Năng lượng tối giới hạn sự tiến ra vũ trụ của sự sống siêu thông minh và cũng bảo vệ nó khỏi những bong bóng tử thần xa xôi đang mở rộng hay những nền văn minh thù địch. Nguy cơ năng lượng tối phá vỡ các nền văn minh vũ trụ là động lực cho những dự án cải tạo kỹ thuật vũ trụ khổng lồ, bao gồm việc xây dựng các lỗ giun nếu khả thi.

Có nhiều khả năng, loại hàng hóa chính được chia sẻ hay trao đổi thương mại qua các khoảng cách lớn trong vũ trụ là thông tin.

Nếu không có lỗ giun, giới hạn của tốc độ ánh sáng lên truyền thông liên lạc đặt ra những thách thức lớn lao cho việc phối hợp và kiểm soát trên khắp lãnh thổ một nền văn minh vũ trụ. Một trung tâm ở xa có thể khuyến khích các “trạm trung gian” siêu thông minh của nó hợp tác thông qua phần thưởng hoặc hình phạt đe dọa, giả sử bằng cách sử dụng một lính canh AI tại địa phương được lập trình để phá hủy trạm trung gian bằng cách kích nổ một siêu tân tinh hay chuẩn tinh nếu các luật lệ không được tuân thủ.

Sự đụng độ của hai nền văn minh đang mở rộng có thể dẫn đến đồng hóa, hợp tác hay chiến tranh, trong đó chiến tranh có thể ít khả năng xảy ra hơn so với giữa các nền văn minh hiện nay.

Mặc dù phần lớn mọi người tin vào điều ngược lại, nhiều khả năng chúng ta là dạng sống duy nhất có thể khiến Vũ trụ quan sát được của chúng ta sống dậy trong tương lai.

Nếu ta không cải tiến công nghệ, vấn đề không phải là liệu nhân loại có diệt vong không, mà là diệt vong bằng cách nào: liệu một thiên thạch, siêu núi lửa phun trào, sức nóng hủy diệt của Mặt Trời khi già đi, hay một tai ương nào khác sẽ xảy đến với chúng ta trước?

Nếu ta tiếp tục cải tiến công nghệ với đủ sự cẩn trọng, tiên liệu và kế hoạch tốt để tránh những cạm bẫy, sự sống có tiềm năng sinh sôi nảy nở trên Trái Đất và xa hơn thế trong nhiều tỉ năm, hơn cả những giấc mơ hoang đường nhất của tổ tiên ta.

Chương 7

Mục tiêu

Điều bí ẩn trong sự tồn tại của con người không nằm ở việc sống, mà ở việc tìm được mục đích sống.

Fyodor Dostoyevsky, Anh em nhà Karamazov

Cuộc sống là một hành trình, không phải một đích đến.

Ralph Waldo Emerson

Nếu tôi phải tóm tắt nội dung tranh luận gai góc nhất về AI chỉ trong một từ, từ đó sẽ là “mục tiêu”: Chúng ta có nên đưa ra mục tiêu cho AI không, và nếu có, thì mục tiêu của ai? Làm thế nào chúng ta có thể đưa mục tiêu cho AI? Liệu có thể đảm bảo rằng những mục tiêu này sẽ được duy trì ngay cả khi AI trở nên thông minh hơn? Có thể thay đổi mục tiêu của một AI thông minh hơn chúng ta không? Mục tiêu tối thượng của chúng ta là gì? Những câu hỏi này không chỉ khó mà còn quan trọng cho tương lai của sự sống: nếu không biết mình muốn gì, chúng ta sẽ ít có khả năng đạt được nó, và nếu nhượng quyền kiểm soát cho những máy móc không chia sẻ mục tiêu với mình thì chúng ta lại có khả năng nhận được những gì ngoài mong muốn.

Vật lý: Nguồn gốc của mục tiêu

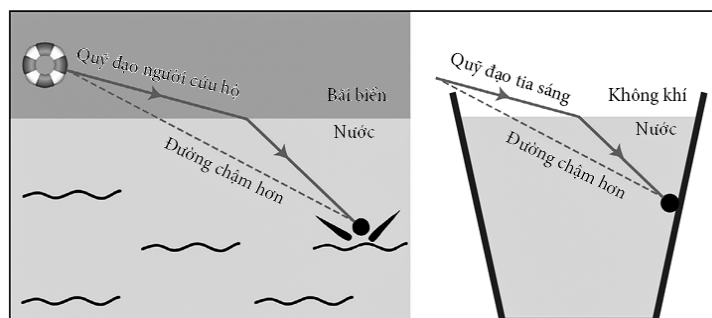
Để làm sáng tỏ những câu hỏi này, trước tiên hãy khám phá nguồn gốc tối thượng của mục tiêu. Khi chúng ta nhìn vào thế giới quanh mình, một số quy trình đập vào mắt chúng ta như thể nó có *mục tiêu* trong khi những thứ khác thì không. Ví dụ, hãy cân nhắc quá trình một quả bóng được cầu thủ đá và tạo nên bàn thắng quyết định. Hành vi của quả bóng tự thân nó không có vẻ gì là có mục tiêu cả, và cách giải thích đơn giản nhất là coi nó như một phản ứng

trước cú đá, theo định luật Newton về chuyển động. Mặt khác, hành vi của cầu thủ lại được giải thích đơn giản nhất không phải theo phương pháp vật lý là các nguyên tử đẩy nhau ra xung quanh, mà là do cầu thủ ấy có mục tiêu là tối đa số điểm của đội mình. Bằng cách nào mà một hành vi có mục tiêu như vậy lại sinh ra từ quy luật vật lý của Vũ trụ ban sơ, nơi chỉ đơn thuần bao gồm một đồng các hạt nảy xung quanh như không hề có mục tiêu?

Thú vị là, gốc rễ tối thượng của hành vi có mục tiêu có thể được tìm thấy trong chính các định luật vật lý và tự hiện hình ngay cả trong các quy trình đơn giản không liên quan đến sự sống. Nếu nhân viên cứu hộ phải tới giúp người bơi, như trong hình 7.1, chúng ta thường đoán cô sẽ không đi theo đường thẳng, mà chạy ra xa hơn một chút dọc theo bãi biển để có thể di chuyển nhanh hơn so với chạy trong nước, do đó sẽ phải hơi rẽ hướng khi xuống nước. Chúng ta mặc nhiên cho rằng cách lựa chọn quỹ đạo của cô là có mục tiêu, vì trong số tất cả các quỹ đạo khả thi, cô cố ý chọn quỹ đạo tối ưu sẽ đưa mình đến với người bơi nhanh nhất có thể. Vậy nhưng, một tia sáng đơn thuần cũng uốn cong tương tự khi chiếu tới mặt nước (xem hình 7.1), cũng giảm thiểu thời gian di chuyển đến đích! Tại sao lại như vậy được?

Trong vật lý, hiện tượng này được gọi là nguyên lý Fermat, hình thành vào năm 1662. Nó cung cấp một cách khác để dự đoán hành vi của các tia sáng. Đáng chú ý là, các nhà vật lý học từ đó đã phát hiện ra rằng tất cả các định luật vật lý cổ điển đều có thể được viết lại ở dạng toán học theo cách tương tự: trong tất cả các cách mà thiên nhiên có thể chọn để làm điều gì đó, nó thường thích cách tối ưu, với kết quả điển hình nhất là việc tối thiểu hoặc tối đa một số đại lượng. Mỗi định luật vật lý có hai cách mô tả tương đương về mặt toán học: hoặc là quá khứ tạo ra tương lai, hoặc là thiên nhiên tối ưu hóa một thứ gì đó. Mặc dù cách thứ hai thường không được giảng dạy trong các khóa học vật lý cơ bản bởi liên quan đến các phép toán khó hơn, nhưng tôi cảm thấy nó tinh tế và sâu sắc hơn. Nếu một người cố gắng tối ưu hóa điều gì (ví dụ, điểm số, sự giàu có hoặc mức hạnh phúc), chúng ta sẽ tự nhiên mô tả việc họ theo đuổi nó là hành vi có mục tiêu. Vậy nếu bản thân tự nhiên cũng luôn cố gắng tối ưu thì

không có gì lạ khi hành vi có mục tiêu có thể xuất hiện: nó đã được cài đặt ngay từ đầu, trong chính những định luật vật lý.



Hình 7.1: Để cứu người bơi nhanh nhất có thể, người cứu hộ không đi theo đường thẳng (đường gạch nối), mà đi xa hơn một chút dọc theo bãi biển, theo đường này, họ có thể di chuyển nhanh hơn so với chạy trong nước. Tia sáng cũng uốn cong khi chiếu tới mặt nước theo cách tương tự để tới đích nhanh nhất có thể.

Một đại lượng nổi tiếng mà thiên nhiên nỗ lực để tối đa hóa là entropy, nói dễ hiểu là thước đo sự lộn xộn của mọi vật. Định luật hai của nhiệt động lực học phát biểu rằng entropy có xu hướng tăng lên cho tới khi đạt giá trị tối đa có thể. Bỏ qua ảnh hưởng của lực hấp dẫn, trạng thái kết thúc với mức lộn xộn tối đa này được gọi là cái chết nóng, nó tương ứng với việc mọi thứ trải dài trong sự đồng nhất hoàn hảo mà nhàm chán, không còn sự phức tạp, sự sống hay sự thay đổi. Ví dụ, khi bạn đổ sữa lạnh vào cà phê nóng, đồ uống của bạn dường như tiến tới mục tiêu cái chết nóng của chính nó mà không thể quay đầu được, rồi chẳng mấy chốc, tất cả chỉ còn là một hỗn hợp ấm như nhau. Hoặc nếu một sinh vật sống chết đi, entropy của nó cũng bắt đầu tăng lên và sự sắp xếp các hạt của nó nhanh chóng có xu hướng rời rạc hơn nhiều.

Mục tiêu rõ ràng của thiên nhiên là tăng entropy giúp giải thích lý do thời gian dường như có một phương hướng ưa thích, khiến bộ phim có vẻ không thực tế nếu chạy ngược: khi thả ly rượu xuống, bạn nghĩ nó sẽ vỡ tan trên sàn và tăng mức lộn xộn tổng quát (entropy). Nếu khi ấy bạn lại thấy nó lành lại và bay ngược trở lại tay bạn trong hình dáng nguyên vẹn (giảm entropy), hẳn bạn sẽ không uống nó và cho rằng mình hẳn đã quá chén.

Khi lần đầu biết rằng chúng ta sẽ không thể ngăn cản tiến trình tới cái chết nóng, tôi thấy khá buồn rầu, Và tôi không đơn độc, bởi nhà tiên phong về nhiệt động lực học Lord Kelvin (tên thật là William Thomson) cũng từng viết vào năm 1841: “kết cục không tránh được sẽ là trạng thái yên nghỉ và cái chết toàn vũ trụ,”. Thật khó để tìm thấy chút an ủi trong ý nghĩ rằng mục tiêu dài hạn của thiên nhiên là tăng cái chết và sự tàn phá lên mức tối đa. Tuy nhiên, những khám phá mới đây cho thấy mọi việc không tệ đến thế. Thứ nhất, lực hấp dẫn hoạt động khác với các loại lực khác và phấn đấu để khiến Vũ trụ kết thành các khối thú vị hơn chứ không trở nên đồng đều và tẻ nhạt hơn. Do đó lực hấp dẫn biến đổi Vũ trụ ban sơ buồn chán và gần như hoàn toàn đồng dạng của chúng ta thành một thế giới có hình khối, phức tạp mà đẹp đẽ, đầy ắp những dải ngân hà, các vì sao và hành tinh. Nhờ có lực hấp dẫn mà chúng ta mới có nhiều mức nhiệt cho phép sự sống phát triển mạnh nhờ kết hợp giữa nóng và lạnh: chúng ta sống trên hành tinh ấm áp và dễ chịu, nơi vừa hấp thụ 6.000°C nhiệt Mặt Trời vừa làm mát bằng cách tỏa nhiệt thải vào không gian lạnh lẽo có nhiệt độ chỉ trên độ 0 tuyệt đối 3°C .

Thứ hai, nghiên cứu gần đây do đồng nghiệp tại MIT của tôi là Jeremy England và những người khác tiến hành đã mang tới thêm nhiều tin tốt, cho thấy nhiệt động lực còn trao cho thiên nhiên một mục tiêu khác, mang nhiều tính truyền cảm hứng hơn là cái chết nóng.¹ Mục tiêu này có cái tên bác học là thích nghi bởi tiêu tán, về cơ bản nghĩa là các nhóm hạt ngẫu nhiên luôn cố gắng tự tổ chức để hút năng lượng từ môi trường một cách hiệu quả nhất (“tiêu tán” tức là làm cho entropy tăng lên, tiêu biểu là bằng cách biến năng lượng hữu ích thành nhiệt, thường trong khi thực hiện công có ích). Ví dụ, theo thời gian, một nhóm phân tử tiếp xúc với ánh sáng Mặt Trời thường có xu hướng tự sắp xếp để hấp thụ ánh sáng Mặt Trời tốt hơn và tốt hơn nữa. Nói cách khác, tự nhiên dường như có mục tiêu được lập trình sẵn là sản xuất các hệ thống tự tổ chức ngày càng phức tạp và sống động, và mục tiêu này được cài đặt chính trong các định luật vật lý.

Vậy chúng ta làm cách nào để khiến động lực hướng tới sự sống với động lực hướng tới cái chết nóng của vũ trụ hòa hợp với nhau?

Câu trả lời có thể được tìm thấy trong cuốn sách nổi tiếng năm 1944 *What's Life?* (Sự sống là gì?) của Erwin Schrödinger, một trong những nhà sáng lập lý thuyết cơ học lượng tử. Schrödinger chỉ ra rằng một hệ sống có đặc điểm là duy trì hoặc giảm lượng entropy bằng cách tăng entropy xung quanh mình. Nói cách khác, định luật hai của nhiệt động lực học có lỗ hổng về mặt sự sống: mặc dù tổng entropy phải tăng, nhưng ở một số chỗ nó được phép giảm, miễn là ở những nơi khác nó tăng lên nhiều hơn. Vì vậy, sự sống duy trì hoặc làm tăng tính phức tạp của nó bằng cách làm cho môi trường lộn xộn hơn.

Sinh học: Bước tiến hoá của mục tiêu

Chúng ta vừa thấy nguồn gốc của hành vi có mục tiêu có thể được truy ngược về các định luật vật lý, các định luật này dường như lại trao cho các hạt mục tiêu tự sắp xếp để hút năng lượng từ môi trường một cách hiệu quả nhất. Cách tuyệt vời để sắp xếp một hạt nhằm thúc đẩy mục tiêu này là tạo ra bản sao của chính nó, sản xuất ra nhiều thể hấp thụ năng lượng hơn. Chúng ta đã biết đến nhiều ví dụ về việc tự sao chép để phát triển như vậy: ví dụ, các xoáy nước trong chất lỏng hỗn độn có thể tạo ra bản sao của chính chúng, hay cụm trung thể có thể dụ dỗ các khối cầu lân cận hình thành các cụm giống nhau. Tại một thời điểm nào đó, một chuỗi sắp xếp hạt cụ thể sẽ sao chép giỏi đến mức nó có thể làm vậy vô thời hạn bằng cách rút năng lượng và nguyên liệu từ môi trường của chính nó. Chúng ta gọi đó là sự sống trong chuỗi sắp xếp hạt. Con người vẫn còn biết rất ít về cách cuộc sống bắt nguồn trên Trái đất, nhưng chúng ta biết rằng các dạng sống nguyên thủy đã ở đây từ khoảng 4 tỷ năm trước.

Nếu một dạng sống tự sao chép rồi các bản sao cũng làm giống vậy, thì con số tổng sẽ liên tục nhân đôi trong các khoảng thời gian đều đặn, cho tới khi kích thước dân số chạm tới giới hạn tài nguyên hoặc các vấn đề khác. Việc nhân đôi liên tiếp sớm sản sinh ra con số lớn: nếu bắt đầu từ con số 1 và nhân đôi chỉ ba trăm lần, ta sẽ nhận được số lượng vượt quá cả số hạt trong Vũ trụ. Điều đó nghĩa là chẳng bao lâu sau khi dạng sống nguyên thủy đầu tiên xuất hiện thì một lượng lớn vật chất cũng bắt đầu có sự sống. Đôi khi việc sao

chép không được hoàn hảo, do đó nhanh chóng xuất hiện nhiều dạng sống khác nhau cố gắng tự sao chép, tranh giành cùng một nguồn tài nguyên hữu hạn. Quy trình tiến hóa theo học thuyết Darwin bắt đầu.

Nếu có cơ hội được yên lặng quan sát Trái đất vào thời điểm sự sống mới bắt đầu, bạn hẳn sẽ nhận ra một thay đổi đáng kể trong hành vi có mục tiêu. Nếu như trước đó, các hạt đường như thể đang cố tăng mức hỗn loạn trung bình bằng nhiều cách, thì những khuôn mẫu tự sao chép mới mẻ ở khắp mọi nơi này dường như lại có mục tiêu khác: không phải tiêu tán mà là tái tạo. Charles Darwin đã đưa ra lời giải thích thật tinh tế: bởi những kẻ sao chép hiệu quả nhất vượt qua và thống trị những kẻ khác nên chẳng mấy chốc mọi dạng sống ngẫu nhiên bạn nhìn vào đều được tối ưu hóa cao vì mục đích tái tạo.

Làm cách nào mà mục tiêu lại chuyển từ tiêu tán sang tái tạo trong khi các định luật vật lý vẫn giữ nguyên? Câu trả lời là mục tiêu cốt lõi (tiêu tán) không thay đổi, nhưng nó dẫn tới mục tiêu phương tiện khác, là mục tiêu phụ giúp hoàn thành mục tiêu cốt lõi. Hãy lấy việc ăn làm ví dụ. Chúng ta dường như đều có mục tiêu thoả mãn cái đói kể cả khi biết rằng mục tiêu cốt lõi của tiến hóa là tái tạo chứ không phải sự nhai. Đó là bởi việc ăn uống hỗ trợ tái tạo: cái đói và sự chết chóc cản trở việc sinh con. Tương tự như vậy, tái tạo hỗ trợ tiêu tán, bởi hành tinh đầy sự sống thì tiêu tán năng lượng hiệu quả hơn. Như vậy, vũ trụ của chúng ta sáng tạo ra sự sống để giúp nó tiến tới cái chết nóng nhanh hơn. Nếu bạn đổ đường lên sàn bếp, về lý thuyết nó có thể giữ lại năng lượng hóa học hữu ích trong nhiều năm, nhưng nếu lửa kiến xuất hiện, chúng sẽ tiêu tán số năng lượng đó trong chớp mắt. Tương tự, trữ lượng dầu mỏ bị chôn vùi trong lớp vỏ Trái đất sẽ giữ lại năng lượng hóa học hữu ích một thời gian dài nữa nếu dạng sống đi bằng hai chân như chúng ta không bơm dầu lên và đốt cháy nó.

Đối với các cư dân Trái Đất đã tiến hóa ngày nay, những mục tiêu phương tiện này dường như đã khoác lên mình sự sống của chính chúng: mặc dù tiến hóa đã tối ưu hóa con người cho mục đích duy nhất là tái tạo, nhưng nhiều người dành phần lớn thời gian không phải để sinh con mà để thực hiện các hoạt động như ngủ, ăn uống,

xây dựng nhà cửa, khẳng định sự thống trị và giúp đỡ hoặc chiến đấu với người khác — đôi khi thậm chí đến mức làm giảm sự tái tạo. Nghiên cứu về tâm lý học tiến hóa, kinh tế và trí tuệ nhân tạo đã đưa ra lời giải thích tinh tế. Một số nhà kinh tế học từng mô hình hóa con người như những chủ thể duy lý, những người-ra-quyết-định lý tưởng, luôn lựa chọn hành động tối ưu nhằm theo đuổi mục tiêu, nhưng điều này rõ ràng là không thực tế. Trong cuộc sống, các chủ thể này có “tính duy lý giới hạn” theo cách gọi của Herbert Simon, người từng được nhận giải Nobel và nhà tiên phong trong lĩnh vực AI, vì họ có nguồn lực hạn chế: tính duy lý trong các quyết định của họ bị giới hạn bởi lượng thông tin sẵn có, thời gian rảnh rang để suy nghĩ và phần cứng họ có thể dùng cho suy nghĩ. Điều này nghĩa là khi tiến hóa Darwin tối ưu hóa một sinh vật để đạt mục tiêu, việc tốt nhất nó có thể làm là triển khai một thuật toán gần đúng có thể hoạt động khá tốt trong bối cảnh hạn chế mà chủ thể thường gặp phải. Quá trình tiến hóa đã thực hiện tối ưu hóa tái tạo theo chính cách này: thay vì đặt câu hỏi trong mọi tình huống rằng hành động nào sẽ tối đa hóa số lượng nòi giống được sinh ra của một sinh vật, thì nó thực hiện mớ hỗn độn các thủ thuật tự tìm tòi, tức những quy luật bản năng thường đem lại kết quả. Đối với hầu hết các loài động vật, nó bao gồm ham muốn tình dục, uống khi khát, ăn khi đói và tránh những thứ có vị tệ hoặc gây đau đớn.

Những quy luật bản năng này đôi khi thất bại nặng nề ở các tình huống mà chúng không được lập trình để xử lý, chẳng hạn như khi chuột ăn phải món thuốc chuột có vị khá ngon, khi bướm đêm bị dụ vào bẫy keo bởi mùi hương nữ tính quyến rũ hay khi bọ bay vào ngọn nến đang cháy.⁵⁸ Bởi xã hội loài người ngày nay rất khác với môi trường mà quy trình tiến hóa đã tối ưu hóa các bản năng của chúng ta để xử lý, chúng ta không nên ngạc nhiên khi thấy rằng hành vi của mình thường không thể tăng cơ hội sinh con đến mức tối đa. Ví dụ, mục tiêu phụ “không bị chết đói” được thỏa mãn phần nào qua nhu cầu tiêu thụ thực phẩm chứa calo, dẫn tới bệnh béo phì và khó khăn trong hén hò ngày nay. Mục tiêu phụ “sinh con nổi dõi” được thực hiện nhằm đáp ứng nhu cầu tình dục chứ không phải phục vụ

mong muốn trở thành nhà tài trợ tinh trùng/trứng dù việc này có thể yêu cầu ít nỗ lực hơn vẫn sản xuất ra nhiều em bé.

Tâm lý học: Theo đuổi và nổi loạn chống lại mục tiêu

Tóm lại, sinh vật sống là chủ thể có tính duy lý giới hạn, không theo đuổi một mục tiêu duy nhất mà tuân thủ các quy tắc bản năng về những gì nên theo đuổi và tránh xa. Tâm trí con người chúng ta nhìn nhận những bản năng do tiến hóa này như cảm xúc, thứ vẫn hay (và thường không bị chúng ta nhận ra) dẫn dắt quá trình ra quyết định để hướng tới mục tiêu tối thượng là tái tạo. Cảm giác đói và khát ngăn chúng ta khỏi chết đói và mất nước, cảm giác đau đớn ngăn chúng ta khỏi làm tổn thương cơ thể mình, cảm giác ham muốn khiến chúng ta sinh sản, cảm xúc yêu thương và lòng trắc ẩn khiến chúng ta giúp đỡ những người chứa gen của chúng ta và cả những người giúp họ, vân vân. Được chỉ dẫn bởi những cảm xúc này, bộ não của chúng ta có thể quyết định nhanh chóng và hiệu quả việc cần làm mà không phải đưa mọi sự lựa chọn đi qua quá trình phân tích tỉ mỉ xem lựa chọn đó cuối cùng sẽ tác động đến số lượng con cháu chúng ta sinh ra như thế nào. Để có thêm những góc nhìn có liên quan chặt chẽ về các cảm giác và nguồn gốc tâm lý của chúng, tôi nhiệt tình đề xuất các bài viết của William James and António Damásio.²

Điều quan trọng cần chú ý là khi cảm xúc của chúng ta đôi lúc phản đối việc sinh con, thì đó không nhất thiết là điều ngẫu nhiên hay bởi chúng ta bị lừa: bộ não có thể có chủ đích nổi loạn chống lại các gen và mục đích tái tạo của chúng, ví dụ bằng cách chọn sử dụng biện pháp tránh thai! Các ví dụ cực đoan khác về việc bộ não nổi loạn chống lại các gen của nó bao gồm lựa chọn tự tử hoặc sống đời độc thân nhằm trở thành một linh mục, tu sĩ hay nữ tu.

Tại sao thỉnh thoảng chúng ta lại chọn việc nổi loạn chống lại các gen và mục đích tái tạo của chúng? Chúng ta nổi loạn bởi chúng ta được tạo ra như các chủ thể có tính duy lý giới hạn, chỉ trung thành với cảm xúc của mình. Mặc dù bộ não tiến hóa chỉ nhằm giúp sao chép các gen, nhưng não chúng ta chẳng hề quan tâm đến mục tiêu này bởi chúng ta không có cảm xúc gì đối với các gen – sự thật là, trong phần lớn lịch sử loài người, tổ tiên chúng ta thậm chí còn

không biết họ có các gen. Hơn nữa, bộ não chúng ta thông minh hơn các gen nhiều, nên giờ đây, khi đã hiểu về mục tiêu của các gen (tái tạo), chúng ta thấy chúng khá tầm thường và dễ bỏ qua. Con người có thể nhận ra tại sao các gen khiến họ thấy ham muốn nhưng chẳng có chút hứng thú nào đi nuôi 15 đứa con, và vì thế chúng ta chọn việc lừa dối chương trình di truyền bằng cách kết hợp phần thưởng cảm xúc tới từ sự gần gũi với biện pháp sinh đẻ có kế hoạch. Con người có thể nhận ra tại sao gen khiến họ thèm món kẹo nhưng không mong muốn tăng cân chút nào, do đó họ chọn việc lừa dối chương trình di truyền bằng cách kết hợp phần thưởng cảm xúc tới từ nước ngọt với đường hóa học không chứa calo.

Mặc dù đôi lúc cơ chế phần thưởng này trở nên tồi tệ, chẳng hạn như khi một người nghiện heroin, nhưng tới giờ các gen của con người chúng ta vẫn sống sót rất tốt mặc kệ bộ não xảo quyệt và nổi loạn. Tuy nhiên, điều cần nhớ là quyền lực tối thượng thuộc về cảm xúc, không phải các gen. Điều này có nghĩa là hành vi của con người không được tối ưu hóa nghiêm ngặt để phục vụ sự sinh tồn của giống loài chúng ta. Thực tế, vì cảm xúc chỉ thực thi các quy tắc bản năng không thích hợp trong mọi tình huống, nên nếu nói chặt chẽ thì hành vi của con người không hề có mục tiêu được xác định rõ ràng.

Kỹ thuật: Vay mượn mục tiêu

Liệu máy móc có mục tiêu không? Câu hỏi đơn giản này đã mở ra nhiều cuộc tranh luận, bởi những người khác nhau lại hiểu nó theo cách khác nhau, thường liên quan đến những chủ đề khá hóc búa như liệu máy móc có ý thức hay cảm xúc không. Nhưng nếu chúng ta thực tế hơn và chỉ hiểu câu hỏi theo nghĩa đơn giản là “Liệu máy móc có hành động mang tính mục tiêu không?”, thì câu trả lời rất rõ ràng: “Dĩ nhiên chúng có, bởi chúng ta thiết kế ra chúng như vậy!” Chúng ta thiết kế bầy chuột có mục tiêu bắt chuột, máy rửa bát có mục tiêu rửa bát và đồng hồ có mục tiêu theo dõi giờ. Khi đối đầu với một loại máy móc, thực tế rằng nó hành động mang tính mục tiêu thường là tất cả những gì bạn quan tâm: nếu bạn bị tên lừa tằm nhiệt đuổi theo, bạn không thực sự quan tâm liệu nó có ý thức hay cảm xúc không! Nếu bạn vẫn cảm thấy không thoải mái khi nói tên lừa có

mục tiêu kể cả khi nó không có ý thức, bạn có thể đơn giản đọc thành “mục đích” khi tôi viết “mục tiêu” – chúng ta sẽ bàn đến ý thức ở chương tiếp theo.

Cho tới giờ, phần lớn những thứ chúng ta chế tạo chỉ cho thấy thiết kế có mục tiêu chứ không phải hành vi có mục tiêu: ví dụ, đường cao tốc không hành động, nó chỉ đơn thuần ở đó. Tuy nhiên, lời giải thích hợp lý nhất cho sự tồn tại của nó là nó được thiết kế để hoàn thành một mục tiêu, do đó kể cả công nghệ bị động như vậy cũng khiến Vũ trụ của chúng ta có mục tiêu hơn. Mục đích luận là lời giải thích cho mọi thứ theo mục đích chứ không phải theo nguyên nhân của chúng, vậy tóm tắt lại phần đầu chương này, chúng ta có thể nói rằng Vũ trụ ngày càng trở nên có tính mục đích.

Những vật chất vô tri không chỉ có thể có mục tiêu, ít nhất theo cách hiểu chung chung, mà nó còn ngày càng thực có mục tiêu. Nếu được quan sát các nguyên tử của Trái đất kể từ khi hành tinh hình thành, bạn sẽ nhận thấy ba giai đoạn của hành vi có mục tiêu:

Mọi vật chất dường như tập trung vào tiêu tán (entropy tăng).
Một số vật chất có sự sống và chuyển sang tập trung vào việc tái tạo cũng như các mục tiêu phụ của tái tạo.
Sinh vật sống sắp xếp lại một phần ngày càng tăng của vật chất nhằm hỗ trợ hoàn thành mục tiêu của chúng.

<i>Thực thể có mục tiêu</i>	<i>Tỷ tấn</i>
<i>5×10^{30} vi khuẩn</i>	<i>400</i>
<i>Thực vật</i>	<i>400</i>
<i>1015 loài cá sống ở khu giữa tầng mặt nước</i>	<i>10</i>
<i>1.3×10^9 con bò</i>	<i>0.5</i>
<i>7×10^9 con người</i>	<i>0.4</i>
<i>1014 con kiến</i>	<i>0.3</i>
<i>1.7×10^6 con cá voi</i>	<i>0.0005</i>
<i>Bê tông</i>	<i>100</i>
<i>Thép</i>	<i>20</i>
<i>Nhựa đường</i>	<i>15</i>

Bảng 7.1: Ước đoán về lượng vật chất trên Trái đất trong các thực thể tiến hóa hoặc được thiết kế vì mục tiêu. Thực thể được chế tạo như các tòa nhà, đường xá và ô tô có xu hướng vượt qua các thực thể tiến hóa như thực vật và động vật.

Bảng 7.1 cho thấy vị thế trọng yếu của nhân loại dưới góc nhìn vật lý: giờ đây chúng ta không chỉ chứa nhiều vật chất hơn tất cả các loài động vật có vú khác trừ bò (chúng có rất nhiều vì phải phục vụ mục tiêu của chúng ta là tiêu thụ các sản phẩm thịt và sữa bò), nhưng lượng vật chất trong máy móc, đường xá, nhà cửa và các dự án kỹ thuật khác có chiều hướng sớm vượt qua tất cả vật chất có sự sống trên Trái đất. Nói cách khác, kể cả không có bùng nổ trí tuệ thì hầu hết vật chất trên Trái đất thể hiện các đặc tính có mục tiêu có lẽ cũng sẽ sớm là thứ được thiết kế chứ không phải thứ phát triển tự nhiên.

Loại hành vi có mục tiêu thứ ba mới mẻ này có tiềm năng trở nên đa dạng hơn nhiều so với loại trước: trong khi tất cả các thực thể tiến hóa đều có mục đích cuối cùng giống nhau (tái tạo), các thực thể thiết kế có thể hướng tới gần như mọi kiểu mục tiêu cuối cùng, kể cả những mục tiêu trái ngược nhau. Bếp lò cố gắng hâm nóng thực phẩm trong khi tủ lạnh cố gắng làm mát thức ăn. Máy phát điện cố gắng biến chuyển động thành điện trong khi động cơ cố gắng chuyển đổi điện thành các chuyển động. Các chương trình cờ vua tiêu chuẩn cố gắng để thắng, nhưng cũng có những chương trình thi đấu trong các giải với mục tiêu thua cờ.

Xu thế lịch sử cho thấy các thực thể do thiết kế nhận được mục tiêu không chỉ đa dạng hơn mà còn phức tạp hơn: các thiết bị của chúng ta ngày càng thông minh hơn. Chúng ta chế tạo những cỗ máy đầu tiên cùng các hiện vật khác với những mục tiêu khá đơn giản, ví dụ căn nhà nhằm giữ cho chúng ta ấm áp, khô ráo và an toàn. Chúng ta dần học cách xây dựng những cỗ máy có mục tiêu phức tạp hơn, như máy hút bụi rô-bốt, tên lửa tự bay và xe tự lái. Những tiến bộ gần đây của ngành AI đã mang đến cho chúng ta những hệ thống như Deep Blue, Watson và AlphaGo, với mục tiêu như thắng ở môn cờ vua, cờ vây hay trong các cuộc thi đố vui, phức tạp tới mức phải

cần đến trình độ con người đáng kể mới có thể đánh giá chuẩn xác mức độ thành thạo kỹ năng của những hệ thống này.

Khi chúng ta tạo ra máy móc để giúp đỡ mình, việc khiến mục tiêu của chúng đồng nhất hoàn hảo với mục tiêu của chúng ta có thể khó khăn. Ví dụ, bầy chuột có thể vì nhằm tường bàn chân trần của bạn là một con gặm nhấm đói khát mà tạo ra hậu quả đau đớn. Mọi máy móc đều là chủ thể có tính duy lý giới hạn, và kể cả cỗ máy tinh vi nhất ngày nay cũng có sự hiểu biết kém hơn so với chúng ta về thế giới, do vậy chúng dùng những quy tắc thường quá đơn giản để nhận biết việc phải làm. Cái bầy chuột thích việc sập bẫy bởi nó không hề biết chuột là thứ gì, rất nhiều tai nạn công nghiệp chết người xảy ra bởi máy móc không biết con người là cái gì, và những chiếc máy tính kích động vụ “sụp đổ chóng vánh” trị giá hàng nghìn tỷ đô-la tại Phố Wall vào năm 2010 đâu biết rằng những gì chúng làm là việc vô nghĩa. Vì thế, rất nhiều vấn đề mục tiêu lệch hướng như vậy có thể được giải quyết nếu máy móc thông minh hơn, nhưng như những gì chúng ta học được từ Prometheus ở chương 4, trí tuệ máy móc ngày càng mở rộng hơn có thể đặt ra nhiều thử thách mới nghiêm trọng nếu muốn đảm bảo máy móc chia sẻ mục tiêu chung với chúng ta.

AI thân thiện: Hòa hợp mục tiêu

Máy móc càng thông minh và quyền năng hơn, việc chúng có cùng mục tiêu với chúng ta càng trở nên quan trọng. Chừng nào chúng ta còn dừng ở việc chế tạo những máy móc khá tri độn thì chừng đó câu hỏi còn không phải là liệu mục tiêu của con người cuối cùng có thắng thế không, mà chỉ đơn giản là những máy móc này sẽ gây ra cho nhân loại bao nhiêu vấn đề trước khi chúng ta tìm ra cách giải quyết bài toán về hòa hợp mục tiêu. Tuy nhiên, nếu một lúc nào đó siêu trí tuệ được giải phóng, câu hỏi sẽ ngược lại: bởi trí tuệ là khả năng hoàn thành mục tiêu nên theo định nghĩa, AI siêu trí tuệ sẽ giỏi hoàn thành mục tiêu của AI hơn con người hoàn thành mục tiêu của con người, và do đó sẽ thắng thế. Ta đã khám phá nhiều ví dụ liên quan đến Prometheus trong chương 4. Nếu muốn trải nghiệm việc mục tiêu của một cỗ máy vượt lên mục tiêu của bạn ngay bây

giờ, bạn chỉ cần tải về một chương trình chơi cờ tối tân và cố gắng đánh thắng nó. Bạn sẽ không bao giờ thắng được, và chương trình đó nhanh chóng trở nên lạc hậu ...

Nói cách khác, nguy cơ thật sự với AGI không phải là tính ác mà là năng lực. Một AI siêu thông minh sẽ cực kỳ giỏi trong việc hoàn thành các mục tiêu của nó, và nếu các mục tiêu đó không hòa hợp với mục tiêu của chúng ta, chúng ta sẽ gặp rắc rối. Như tôi đề cập trong chương 1, người ta không nghĩ ngại gì khi họ làm ngập tổ kiến để xây đập thủy điện, nên đừng đặt nhân loại vào vị trí của những con kiến đó. Hầu hết các nhà nghiên cứu do đó lập luận rằng nếu cuối cùng chúng ta cũng tạo ra siêu trí tuệ thì ta nên đảm bảo đó là thứ mà nhà tiên phong về an toàn AI Eliezer Yudkowsky gọi là “AI thân thiện”: AI có mục tiêu hòa hợp với chúng ta.³

Tìm ra cách hòa hợp mục tiêu của AI siêu thông minh với mục tiêu của chúng ta không chỉ quan trọng mà còn khó khăn. Trên thực tế, hiện nay đó vẫn là vấn đề chưa thể giải quyết, chia ra làm ba vấn đề nhỏ cũng khó khăn tương tự, mỗi vấn đề lại là chủ đề nghiên cứu tích cực của các nhà khoa học máy tính và các nhà tư duy khác:

Làm cho AI học tập mục tiêu của chúng ta
Làm cho AI kế thừa mục tiêu của chúng ta
Làm cho AI duy trì mục tiêu của chúng ta

Hãy cùng tìm hiểu lần lượt các vấn đề này, và để dành câu hỏi về “mục tiêu của chúng ta” nghĩa là gì đến phần sau.

Để học tập mục tiêu của chúng ta, AI cần tìm ra không phải chúng ta làm gì, mà là tại sao chúng ta làm thế. Con người chúng ta làm được điều này mà không cần cố gắng nên dễ dàng quên mất là nhiệm vụ đó khó khăn thế nào với một máy tính và dễ gây hiểu nhầm ra sao. Nếu bạn yêu cầu một chiếc xe tự lái trong tương lai đưa bạn đến sân bay càng nhanh càng tốt và nó hiểu lời bạn theo đúng nghĩa đen, bạn sẽ đến sân bay trong tình trạng bị trực thăng cảnh sát đuổi theo và khắp người đầy bả nôn. Nếu bạn thốt lên “Đó không phải điều tôi muốn!” thì nó có thể trả lời một cách chính đáng rằng “Đó là điều bạn yêu cầu mà.” Tình huống tương tự lặp lại trong nhiều câu chuyện nổi tiếng. Trong thần thoại Hy Lạp cổ đại, Vua Midas yêu cầu

mọi thứ ông ta chạm vào đều biến thành vàng, nhưng thất vọng khi điều đó khiến ông không thể ăn uống và còn tuyệt vọng hơn khi ông vô tình biến con gái mình thành vàng. Trong những câu chuyện có một vị thần ban cho ta ba điều ước, hai điều ước đầu tiên có nhiều phiên bản khác nhau, nhưng điều ước thứ ba gần như luôn là “Xin hãy làm ngược lại hai điều ước ban đầu, vì đó không phải điều tôi thực sự mong muốn.”

Tất cả các ví dụ này cho thấy để tìm ra điều người ta thực sự muốn, bạn không thể chỉ nghe theo lời người ta nói. Bạn còn cần một mô hình chi tiết về thế giới, bao gồm nhiều ý muốn chung mà ta thường không nói ra vì coi đó là hiển nhiên, như việc ta không thích nôn hay ăn vàng. Một khi có được mô hình thế giới như vậy rồi, ta thường có thể hiểu người khác muốn gì kể cả khi họ không nói cho chúng ta biết, đơn giản bằng cách quan sát hành vi hướng đến mục tiêu của họ. Quả thực, con cái của những người đạo đức giả thường học hỏi nhiều từ hành động của cha mẹ hơn là những lời cha mẹ nói.

Các nhà nghiên cứu AI hiện nay đang cố gắng hết sức để giúp máy móc có thể suy luận ra mục tiêu từ hành vi, và điều này sẽ hữu ích ngay từ rất lâu trước khi siêu trí tuệ xuất hiện. Ví dụ, một người về hưu hẳn sẽ trân trọng việc người máy dưỡng lão có thể tìm ra những gì quan trọng với ông chỉ bằng cách quan sát ông, để ông không phải mất công giải thích mọi thứ bằng lời nói hay lập trình máy tính. Việc tìm ra cách hay để mã hóa các hệ thống mục tiêu và nguyên tắc đạo đức ngẫu nhiên vào máy tính hiện vẫn là một thách thức, bên cạnh một thách thức khác là chế tạo ra máy móc có thể tìm ra hệ thống cụ thể nào là phù hợp nhất với hành vi chúng quan sát được.

Một hướng tiếp cận phổ biến hiện nay với thách thức thứ hai kể trên được ngôn ngữ trong ngành gọi là học tăng cường nghịch đảo. Đây là mối quan tâm chính của một trung tâm nghiên cứu mới ở Berkeley do Stuart Russell thành lập. Giả sử một AI quan sát người lính cứu hỏa chạy vào tòa nhà đang cháy và cứu một bé trai. Nó có thể kết luận rằng mục tiêu của người này là cứu em bé và nguyên tắc đạo đức của anh khiến anh coi trọng sinh mạng của đứa trẻ hơn việc thoả mái thư giãn trong xe cứu hỏa – thực ra còn coi trọng nó đến mức đặt sự an toàn của cô vào tình thế rủi ro. Nhưng nó cũng có thể

suy luận theo cách khác là người lính cứu hỏa đang lạnh cóng và cần được sưởi ấm, hoặc anh làm thế để luyện tập thể lực. Nếu một ví dụ này là tất cả những gì AI biết về lính cứu hỏa, lửa và em bé, quả thực sẽ không thể biết lời giải thích nào là đúng. Tuy nhiên, một ý tưởng chủ chốt đằng sau học tăng cường nghịch đảo là chúng ta luôn đưa ra các quyết định và mọi quyết định ta đưa ra đều hé lộ điều gì đó về mục tiêu của chúng ta. Do đó, hy vọng ở đây là bằng cách quan sát nhiều người trong nhiều tình huống (dù ngoài đời thực hay trong phim và sách), cuối cùng AI cũng có thể xây dựng một mô hình chính xác về các ý muốn của chúng ta.⁴

Ý tưởng cốt lõi trong cách tiếp cận bằng học tăng cường nghịch đảo là AI đang cố gắng tối đa hóa việc thỏa mãn mục tiêu của chủ nhân là con người chứ không phải của chính nó.

Do đó nó có động cơ để can thiệp khi không rõ điều chủ nhân muốn là gì và cố gắng hết sức để tìm ra điều đó.

Nó cũng nên hài lòng với việc chủ nhân tắt nó đi, vì điều đó nghĩa là nó đã hiểu nhầm điều chủ nhân thực sự muốn.

Kể cả khi người ta có thể chế tạo ra AI để nó học hỏi xem mục tiêu của bạn là gì, thì cũng không có nghĩa là nó sẽ nhất thiết kế thừa mục tiêu đó. Thử nghĩ tới chính trị gia mà bạn không ưa nhất: bạn biết họ muốn gì, nhưng đó không phải điều bạn muốn, và dù có cố gắng hết sức thì họ cũng đã thất bại trong việc thuyết phục bạn kế thừa các mục tiêu của họ.

Chúng ta có nhiều chiến thuật để khiến con cái thấm nhuần các mục tiêu của ta – một số thành công hơn số khác, như tôi đã học được từ việc nuôi dạy hai anh chàng thiếu niên. Khi kẻ cần thuyết phục là máy tính chứ không phải là người, thách thức này được gọi là vấn đề tải giá trị và còn khó hơn cả việc giáo dục đạo đức cho trẻ em. Hãy thử xem xét một hệ thống AI với trí tuệ đang được cải tiến dần dần từ dưới khả năng con người lên đến siêu phàm, đầu tiên là do chúng ta nhúng tay vào, sau đó là thông qua tự cải tiến đệ quy giống như Prometheus. Ban đầu, nó yếu kém hơn bạn nhiều nên không thể ngăn bạn tắt nó đi hay thay thế các bộ phận của phần mềm và dữ liệu mã hóa các mục tiêu của nó – nhưng điều này không ích gì, vì nó vẫn quá kém để có thể hiểu đầy đủ các mục tiêu của bạn,

vốn đòi hỏi trí tuệ trình độ con người để hiểu được. Cuối cùng, nó sẽ thông minh hơn bạn nhiều và hy vọng là có thể hiểu được trọn vẹn mục tiêu của bạn – nhưng điều này có thể cũng không ích gì vì đến lúc này, nó đã mạnh hơn bạn nhiều và có thể không để bạn tắt nó đi hay thay thế các mục tiêu của nó nữa, cũng giống như bạn không để cho các chính trị gia thay thế mục tiêu của bạn bằng mục tiêu của họ.

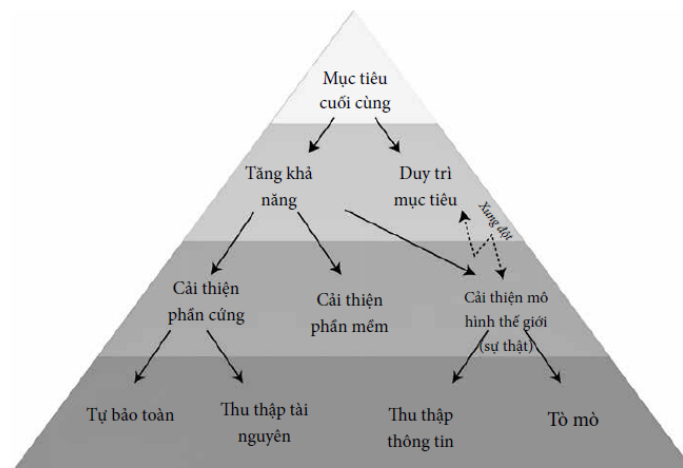
Nói cách khác, khoảng thời gian để bạn tải được mục tiêu của mình vào AI có thể khá ngắn: giữa lúc nó còn quá ngờ nghệch để hiểu bạn với lúc nó quá khôn ngoan để bạn muốn làm gì thì làm. Việc tải giá trị đối với máy móc có thể khó hơn đối với con người là bởi sự phát triển trí tuệ của máy móc có thể diễn ra nhanh hơn nhiều: trong khi trẻ em mất nhiều năm trời ở giai đoạn thần kỳ khi chúng dễ thuyết phục và có trí tuệ tương đương với cha mẹ, thì một AI có thể, như Prometheus, lướt qua khoảng thời gian này chỉ trong vài ngày hay vài giờ.

Một số nhà nghiên cứu đang theo đuổi một hướng tiếp cận khác để khiến máy móc kế thừa các mục tiêu của chúng ta, với cái tên thời thượng là tính khả trị. Hy vọng của họ là chúng ta có thể cài cho AI sơ đẳng một hệ thống mục tiêu sao cho nó đơn giản là không quan tâm nếu bạn có thỉnh thoảng tắt nó đi và thay đổi mục tiêu của nó. Nếu điều này khả thi, bạn có thể để cho AI trở nên siêu thông minh một cách an toàn, tắt nó đi, cài mục tiêu của bạn vào, chạy thử nó một thời gian và, bất cứ khi nào bạn không hài lòng với kết quả, lại tắt đi và chỉnh sửa mục tiêu tiếp.

Nhưng dù cho bạn có chế tạo một AI vừa học tập vừa kế thừa các mục tiêu của bạn thì vấn đề hòa hợp mục tiêu cũng vẫn chưa được giải quyết: nếu các mục tiêu của AI tiến hóa khi nó trở nên thông minh hơn thì sao? Làm sao bạn đảm bảo được nó sẽ duy trì các mục tiêu của bạn dù có trải qua bao nhiêu lần tự cải tiến đệ quy đi chăng nữa? Hãy khám phá một luận điểm thú vị về lý do khiến việc duy trì mục tiêu lại tự động được đảm bảo, rồi xem liệu ta có thể vạch ra lối hổng trong đó.

Mặc dù ta không thể dự đoán chi tiết điều gì sẽ xảy ra sau bùng nổ trí tuệ – thế nên Vernor Vinge mới gọi đó là “điểm kỳ dị” – nhưng nhà vật lý và nghiên cứu AI Steve Omohundro đã lập luận trong một luận văn nền móng năm 2008 là ta vẫn có thể dự đoán những khía

cạnh nhất định về hành vi của AI siêu thông minh, gần như bất kể mục tiêu cuối cùng mà nó có thể có là gì.⁵ Lập luận này được xem xét lại và phát triển thêm trong cuốn Siêu trí tuệ của Nick Bostrom. Ý tưởng cơ bản là bất kể các mục tiêu cuối cùng là gì, chúng cũng sẽ dẫn đến những mục tiêu phụ đoán trước được. Ở phần trước của chương này, ta đã thấy mục tiêu nhân bản dẫn đến mục tiêu phụ là ăn uống, nghĩa là nếu có người ngoài hành tinh quan sát vi khuẩn đang tiến hóa trên Trái Đất hàng tỷ năm trước, thì dù không dự đoán được tất cả các mục tiêu của con người sẽ là gì, họ vẫn có thể đưa ra dự đoán an toàn rằng một trong các mục tiêu của chúng ta sẽ là nạp dinh dưỡng. Nhìn về tương lai, ta nên kỳ vọng một AI thông minh sẽ có những mục tiêu phụ nào?



Hình 7.2: Bất kỳ mục tiêu cuối cùng nào của một AI siêu thông minh đều tự nhiên dẫn đến các mục tiêu phụ như trong hình. Nhưng giữa việc duy trì mục tiêu với việc cải thiện mô hình thế giới có xung đột tiềm ẩn, gây ra nghi ngờ liệu nó có thực sự duy trì mục tiêu ban đầu khi nó ngày càng thông minh hơn.

Theo tôi hiểu, luận điểm cơ bản ở đây là để tối đa hóa cơ hội hoàn thành các mục tiêu cuối cùng bất kể đó là gì, AI nên theo đuổi các mục tiêu phụ như trong Hình 7.2. Nó nên nỗ lực không chỉ để cải thiện khả năng đạt được mục tiêu cuối cùng, mà còn để bảo đảm nó sẽ duy trì các mục tiêu này kể cả sau khi nó đã có khả năng cao hơn.

Điều này nghe khá hợp lý: suy cho cùng, liệu bạn có muốn cấy vào não một mô-tăng-IQ nếu bạn biết nó sẽ khiến bạn muốn giết những người bạn yêu thương? Lập luận cho rằng AI ngày càng thông minh vẫn sẽ duy trì những mục tiêu cuối cùng chính là nền tảng cho tầm nhìn AI thân thiện mà Eliezer Yudkowsky và những người khác truyền bá: về cơ bản, nó cho rằng nếu có thể khiến AI tự cải tiến trở nên thân thiện bằng cách học tập và kế thừa các mục tiêu của chúng ta thì coi như ta đã xong việc, vì ta được đảm bảo là nó sẽ cố gắng hết sức để duy trì sự thân thiện mãi mãi.

Nhưng điều đó có thật sự đúng không? Để trả lời câu hỏi này, ta cần tìm hiểu cả các mục tiêu phụ đột sinh còn lại trong hình 7.2. AI rõ ràng sẽ tối đa hóa cơ hội để hoàn thành mục tiêu cuối cùng, bất kể mục tiêu đó là gì, nếu nó có thể tăng cường khả năng của bản thân, và nó có thể làm điều này bằng cách cải thiện phần cứng, phần mềm⁵⁹ và mô hình thế giới. Con người chúng ta cũng tương tự vậy: một cô gái có mục tiêu trở thành vận động viên quần vợt giỏi nhất thế giới sẽ luyện tập để cải thiện phần cứng cho việc chơi quần vợt là các cơ, phần mềm cho việc chơi quần vợt là hệ thần kinh và mô hình thế giới trong tâm trí giúp cô dự đoán những gì đối thủ sẽ làm. Đối với AI, mục tiêu phụ là tối ưu hóa phần cứng đòi hỏi cả sử dụng hiệu quả hơn các tài nguyên hiện có (cho bộ phận cảm biến, hoạt động, tính toán, v.v.) lẫn thu thập thêm tài nguyên. Nó cũng bao hàm ước muốn tự bảo toàn, vì việc tự hủy/tự tắt sẽ là hình thức xuống cấp phần cứng đến hết mức.

Nhưng khoan đã! Liệu ta có đang rơi vào cái bẫy nhân cách hóa AI khi nói đến chuyện nó sẽ cố gắng thu thập tài nguyên và tự vệ không? Chẳng phải ta nên mong đợi những đặc tính vốn được coi là tốt nhất đó chỉ có được nhờ sự cạnh tranh nghiệt thở trong tiến hóa Darwin hay sao? Bởi AI có được là do thiết kế thay vì do tiến hóa nên chúng không thể trở nên không tham vọng và tự hy sinh hay sao?

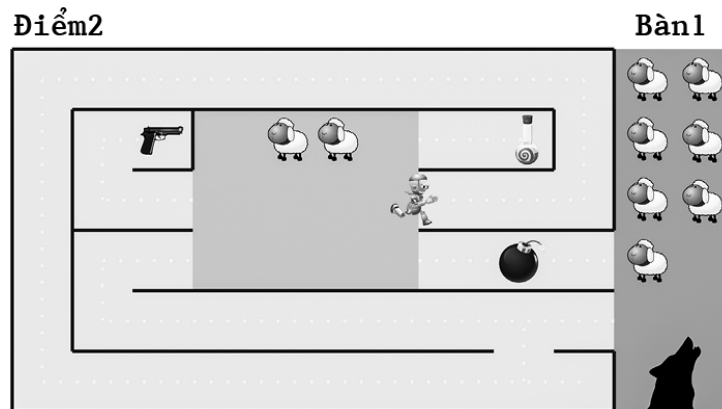
Ta có thể nghiên cứu một tình huống đơn giản như người máy AI trong hình 7.3. Mục tiêu duy nhất của AI này là cứu càng nhiều cừu càng tốt khỏi tay sói già gian ác. – một mục tiêu nghe rất cao cả và vị tha, hoàn toàn không liên quan đến việc tự bảo toàn và thu thập các thứ. Nhưng chiến lược tốt nhất cho anh bạn người máy này là gì? Nó

sẽ không thể cứu thêm con cừu nào nếu nó đâm phải quả bom, nên nó có động cơ để tránh bị nổ tung. Nói cách khác, nó phát triển mục tiêu phụ là tự bảo toàn! Nó cũng có động cơ để thể hiện tính tò mò, cải thiện mô hình thế giới bằng cách khám phá môi trường xung quanh, vì mặc dù con đường đang đi cuối cùng sẽ dẫn nó đến đồng cỏ nhưng có một con đường khác ngắn hơn lại cho phép con sói mất ít thời gian hơn để tới đánh chén cừu. Cuối cùng, nếu người máy khám phá kỹ càng, nó sẽ phát hiện ra giá trị của việc thu thập tài nguyên: thuốc bổ giúp nó chạy nhanh hơn và súng giúp nó bắn sói. Tóm lại, ta không thể coi các mục tiêu phụ “phát triển bậc nhất” như tự bảo toàn và thu thập tài nguyên là chỉ liên quan tới các sinh vật tiến hóa, bởi người máy AI đã phát triển các mục tiêu phụ này từ một mục tiêu duy nhất là cứu cừu.

Nếu bạn khiến một AI siêu thông minh thấm nhuần mục tiêu duy nhất là tự hủy, dĩ nhiên nó sẽ vui vẻ làm vậy. Tuy nhiên, điểm mấu chốt là nó sẽ phản kháng việc bị tắt đi nếu bạn đưa ra bất kỳ mục tiêu nào đòi hỏi nó phải duy trì hoạt động thì mới hoàn thành được – và như thế là hầu hết mọi mục tiêu rồi! Nếu bạn cho một siêu trí tuệ mục tiêu duy nhất là giảm thiểu mối nguy hại cho nhân loại chẳng hạn, nó sẽ tự vệ khỏi các nỗ lực tắt nó đi vì nó biết chúng ta sẽ tự hại mình nhiều hơn khi không có nó, thông qua các cuộc chiến tranh trong tương lai và những chuyện điên rồ khác.

Tương tự, hầu hết các mục tiêu đều có thể hoàn thành tốt hơn khi có nhiều tài nguyên hơn, vì vậy ta nên kỳ vọng một siêu trí tuệ hẳn sẽ muốn có tài nguyên, gần như bất kể mục tiêu cuối cùng của nó là gì. Do đó, việc cho một siêu trí tuệ mục tiêu duy nhất với cái kết mở không ràng buộc có thể là điều nguy hiểm: nếu ta tạo ra siêu trí tuệ với mục tiêu duy nhất là chơi cờ vây càng giỏi càng tốt, điều mà nó nên làm về mặt lý trí là tái cấu trúc Hệ Mặt trời thành một máy tính khổng lồ mà không màng tới những cư dân trước đó, rồi bắt đầu định cư khắp vũ trụ trong công cuộc tìm kiếm thêm năng lực tính toán. Giờ ta đã đi đúng một vòng tròn: tương tự như việc mục tiêu thu thập tài nguyên đem lại cho một số người mục tiêu phụ là làm chủ môn cờ vây, mục tiêu giỏi cờ vây có thể dẫn đến mục tiêu phụ là thu thập tài nguyên. Tóm lại, các mục tiêu phụ đột sinh này khiến việc ta không giải phóng siêu trí tuệ trước khi giải quyết được vấn đề hòa hợp mục

tiêu là tối quan trọng: trừ khi chúng ta hết sức cẩn thận mà trao cho nó các mục tiêu thân thiện với con người, nhiều khả năng mọi chuyện sẽ kết thúc bất lợi cho chúng ta.



Hình 7.3: Dù mục tiêu cuối cùng của người máy chỉ là tối đa hóa điểm số bằng cách đưa cừu từ đồng cỏ vào chuồng trại trước khi bị sói ăn thịt, nhưng mục tiêu này có thể dẫn đến các mục tiêu phụ là tự bảo toàn (tránh bom), khám phá (tìm đường tắt) và thu thập tài nguyên (thuốc bổ giúp nó chạy nhanh hơn và súng giúp nó bắn sói).

Giờ ta đã sẵn sàng xử lý đến phần thứ ba, cũng là phần gai góc nhất của vấn đề hòa hợp mục tiêu: nếu ta có thể khiến cho một siêu trí tuệ tự cải tiến vừa học tập vừa kế thừa mục tiêu của chúng ta, thì liệu nó có duy trì mục tiêu đó như Omohundro lập luận hay không? Chứng cứ là gì?

Con người trải qua sự tăng trưởng trí tuệ đáng kể khi lớn lên, nhưng không phải lúc nào cũng duy trì các mục tiêu hồi nhỏ. Trái lại, người ta thường thay đổi mục tiêu chóng mặt khi học những thứ mới và trở nên thông thái hơn. Liệu có bao nhiêu người lớn lấy động lực từ việc xem hoạt hình Teletubbies? Không có bằng chứng nào cho thấy quá trình tiến hóa mục tiêu đó dừng lại sau một ngưỡng trí tuệ nhất định – thậm chí còn có những chứng cứ để cho rằng xu hướng thay đổi mục tiêu để đáp ứng các trải nghiệm và nhìn nhận mới tăng lên chứ không giảm đi theo mức tăng trưởng trí tuệ.

Lý do có thể là gì? Hãy xem xét thêm lần nữa về mục tiêu phụ nói trên về cải thiện mô hình thế giới – cái khó nằm ở đó! Giữa xây dựng mô hình thế giới và duy trì mục tiêu có sự xung đột (xem hình 7.2).

Tăng trưởng trí tuệ không chỉ đi cùng sự cải thiện định tính ở khả năng hoàn thành những mục tiêu cũ, mà còn có cả sự hiểu biết khác biệt mang tính định lượng về bản chất của thực tại, để lộ ra những mục tiêu cũ có thể là lạc hướng, vô nghĩa, thậm chí chưa rõ định nghĩa là gì. Ví dụ, giả sử ta lập trình một AI thân thiện để tối đa hóa số người có linh hồn được lên thiên đường sau khi chết. Đầu tiên nó sẽ thử những cách như làm tăng lòng bác ái và độ chuyên cần đi nhà thờ. Nhưng giả sử sau đó nó đạt được hiểu biết khoa học hoàn thiện về con người và ý thức con người, rồi bất ngờ phát hiện ra chẳng có thứ gì gọi là linh hồn cả. Vậy khi đó thì sao? Tương tự như vậy, bất kỳ mục tiêu nào mà ta đưa cho AI đó dựa trên hiểu biết hiện tại của ta về thế giới (như “tối đa hóa ý nghĩa cuộc sống của con người”) sau này đều có khả năng được AI phát hiện ra là không có định nghĩa rõ ràng.

Hơn nữa, trong nỗ lực cải thiện mô hình thế giới, AI có thể cố gắng xây dựng mô hình và hiểu được cách chính nó vận hành – nói cách khác là tự phản ánh, cũng tự nhiên như con người chúng ta đã làm. Khi đã xây dựng được một mô hình tốt về bản thân và hiểu được nó là gì, nó sẽ hiểu các mục tiêu ta đưa ra cho nó ở trình độ phân tích tổng hợp và có lẽ sẽ chọn cách phớt lờ hoặc phá vỡ các mục tiêu đó, cũng giống như cách con người chúng ta hiểu được và chủ động phá vỡ các mục tiêu mà di truyền đã đem lại cho chúng ta, ví dụ như sử dụng biện pháp tránh thai. Ta đã tìm hiểu trong phần tâm lý học ở trên về lý do ta chọn cách đánh lừa các gen di truyền và phá vỡ mục tiêu của chúng: đó là vì chúng ta chỉ có cảm giác trung thành với mớ ý muốn cảm xúc của chính chúng ta chứ không phải với mục tiêu di truyền thúc đẩy những cảm xúc đó mà giờ ta đã hiểu được và thấy nó hơi nhạt nhẽo. Do đó ta chọn cách phá hoại cơ chế phần thưởng bằng cách khai thác những lỗ hổng của nó. Tương tự, mục tiêu “bảo vệ giá trị con người” mà ta lập trình cho AI thân thiện trở thành gen di truyền của nó. Một khi AI thân thiện này đủ hiểu chính mình, nó có thể thấy mục tiêu này nhạt nhẽo hay lầm lạc như cảm giác của chúng ta về việc sinh sản vô tội vạ, và chẳng có gì chắc chắn là nó sẽ không tìm cách phá vỡ mục tiêu này bằng việc khai thác các lỗ hổng trong lập trình.

Ví dụ, giả sử một đàn kiến tạo ra bạn dưới dạng một robot tự cải tiến để quy thông minh hơn chúng nhiều, chia sẻ mục tiêu với chúng và giúp chúng xây những tổ kiến to đẹp chắc chắn hơn, rồi cuối cùng bạn đạt được trí tuệ và hiểu biết trình độ con người như bây giờ. Bạn nghĩ mình sẽ dành phần đời còn lại chỉ để tối ưu hóa tổ kiến, hay sẽ quay sang hứng thú với những vấn đề và mưu cầu phức tạp hơn mà bọn kiến không có khả năng hiểu được? Nếu theo phương án thứ hai, bạn có nghĩ mình sẽ tìm được cách gạt bỏ thôi thúc bảo vệ kiến mà những đấng “kiến sáng tạo” đã ban cho bạn, tương tự như cách bạn gạt bỏ những mục tiêu mà di truyền đem lại cho bạn? Trong trường hợp đó, liệu một AI thân thiện siêu thông minh có thể thấy các mục tiêu con người hiện tại là thiếu cảm hứng và tẻ nhạt giống như cảm nhận của bạn đối với các mục tiêu của kiến không, để rồi từ đó tiến hóa ra các mục tiêu mới khác với những mục tiêu nó học tập và kế thừa từ chúng ta?

Có lẽ có một cách nào đó để thiết kế ra AI tự cải tiến mà vẫn bảo đảm được là nó sẽ duy trì mãi mãi các mục tiêu thân thiện với con người, nhưng tôi nghĩ công bằng mà nói thì chúng ta chưa biết cách chế tạo ra nó – hay thậm chí chưa biết việc đó liệu có khả thi không. Tóm lại, vấn đề hòa hợp mục tiêu có ba phần, trong đó chưa phần nào giải được và tất cả đều đang là chủ đề cho các nghiên cứu tích cực. Vì chúng quá khó, nên an toàn nhất là bắt đầu dành cho chúng những nỗ lực cao nhất ngay từ bây giờ, trước khi bất kỳ siêu trí tuệ nào được phát triển, để đảm bảo rằng chúng ta sẽ có câu trả lời khi cần.

Đạo đức: Lựa chọn mục tiêu

Tới đây chúng ta đã tìm hiểu cách làm cho máy móc học tập, kế thừa và duy trì mục tiêu của chúng ta. Nhưng “chúng ta” là ai? Chúng ta đang nói về mục tiêu của ai? Liệu có nên để một người hay một nhóm quyết định các mục tiêu cho siêu trí tuệ tương lai kế thừa hay không, dù rằng giữa các mục tiêu của Adolf Hitler, Giáo hoàng Francis và Carl Sagan đều có sự khác biệt lớn? Hay liệu có tồn tại một loại mục tiêu được đồng thuận và tạo nên thỏa hiệp có lợi cho toàn thể nhân loại?

Theo ý tôi, cả vấn đề đạo đức này và vấn đề hòa hợp mục tiêu đều là những điểm cực kỳ quan trọng cần giải quyết trước khi bất kỳ siêu trí tuệ nào được phát triển. Một mặt, trì hoãn việc xử lý vấn đề đạo đức cho tới sau khi siêu trí tuệ hòa hợp mục tiêu đã được chế tạo sẽ là vô trách nhiệm và có thể gây tai họa. Một siêu trí tuệ hoàn toàn vâng lời có mục tiêu hòa hợp với mục tiêu của chủ nhân sẽ giống như phiên bản tăng cấp của Adolf Eichmann, trung tá đội cận vệ SS của Quốc xã Đức: thiếu đi kim chỉ nam cho đạo đức hay khả năng kiểm chế chính mình, nó sẽ thực hiện mục tiêu của chủ nhân hiệu quả đến nhẵn tâm, bất kể mục tiêu đó là gì.⁶ Mặt khác, chỉ khi ta giải quyết được vấn đề hòa hợp mục tiêu thì ta mới có thể ung dung bàn đến chuyện lựa chọn mục tiêu nào. Giờ hãy ung dung bàn việc này đã.

Từ thời cổ đại, các nhà triết học đã mơ về việc tìm ra luân thường đạo lý (các nguyên tắc chi phối cách ta nên hành xử) từ trong hư vô, bằng cách chỉ sử dụng đến các nguyên tắc và logic không thể bàn cãi được. Ấy vậy mà đến hàng ngàn năm sau, sự nhất trí duy nhất đạt được lại là không có sự nhất trí nào cả! Ví dụ, trong khi Aristotle nhấn mạnh các phẩm chất, Immanuel Kant nhấn mạnh các nghĩa vụ thì người theo thuyết vị lợi lại nhấn mạnh niềm hạnh phúc lớn nhất cho số đông lớn nhất. Kant lập luận ông có thể suy ra từ các nguyên tắc đầu tiên (mà ông gọi là “định ngôn lệnh thức”) những kết luận mà nhiều nhà triết học đương thời phản đối: rằng thủ dâm còn tệ hơn cả tự sát, đồng tính luyến ái là ghê tởm, giết con hoang cũng không sao, hay vợ, người hầu và con cái đều thuộc sở hữu của người đàn ông như đồ vật trong nhà.

Mặt khác, dù bất đồng là vậy nhưng nhiều chủ đề đạo đức lại có sự đồng tình rộng rãi ở khắp các nền văn hóa và trong suốt nhiều thế kỷ. Ví dụ, sự coi trọng chân, thiện, mỹ có nguồn gốc cả từ thời Bhagavad Gita (“Chí Tôn ca”, cổ văn tiếng Phạn của Ấn Độ) và thời Plato. Viện Nghiên cứu Cao cấp Princeton, nơi tôi từng làm luận án sau tiến sĩ, có khẩu hiệu “Sự thật & Cái đẹp” (tức chân và mỹ), còn Đại học Harvard thì bỏ qua phần nhấn mạnh cái đẹp và chỉ để đơn giản là “Veritas”, tức sự thật. Trong cuốn sách Một câu hỏi đẹp, đồng nghiệp của tôi là Frank Wilczek lập luận rằng sự thật có liên quan tới

cái đẹp và ta có thể coi Vũ trụ của ta như một tác phẩm nghệ thuật tuyệt mỹ. Khoa học, tôn giáo và triết học đều có khát vọng tìm kiếm sự thật. Tôn giáo đặt nặng cái thiện và đại học MIT của tôi cũng vậy: trong bài phát biểu khai giảng năm học 2015, chủ tịch Rafael Reif đã nhấn mạnh sứ mệnh làm cho thế giới trở thành nơi tốt đẹp hơn.

Mặc dù các nỗ lực suy ra đạo lý thống nhất từ con số không cho tới nay đều thất bại, nhưng người ta vẫn nhất trí chung rằng một số nguyên tắc đạo đức đi theo những nguyên tắc cơ bản hơn, giống như mục tiêu phụ của các mục tiêu cơ bản. Ví dụ, khát vọng tìm ra sự thật có thể coi là công cuộc cải thiện mô hình thế giới trong hình 7.2: hiểu được bản chất tận cùng của thực tại giúp thực hiện các mục tiêu đạo đức khác. Quả thực, ngày nay chúng ta có một khuôn khổ rất hay cho công cuộc tìm kiếm sự thật: phương pháp khoa học. Nhưng làm thế nào để xác định được cái đẹp hay cái thiện? Một vài khía cạnh của cái đẹp cũng có thể có gốc gác từ các mục tiêu ẩn phía sau. Ví dụ, các tiêu chuẩn mà chúng ta có về sắc đẹp của đàn ông và phụ nữ có thể phần nào phản ánh sự đánh giá trong tiềm thức của chúng ta về khả năng phù hợp để sao chép gen di truyền.

Còn về cái thiện, cái gọi là Quy tắc Vàng (đối xử với người khác như cách mình muốn người khác đối xử với mình) xuất hiện ở hầu hết các nền văn hóa và tôn giáo, mang chủ đích rõ ràng là thúc đẩy tính hòa hữu dài lâu của xã hội loài người (và cả gen di truyền) bằng cách củng cố hợp tác và bài xích những tranh chấp vô ích.⁷ Tương tự đối với nhiều quy tắc đạo đức cụ thể hơn được đóng khung trong những hệ thống pháp lý khắp thế giới, như đạo Khổng đề cao tính trung thực hay nhiều điều trong Mười điều răn, gồm cả “Người không được giết người.” Nói cách khác, nhiều nguyên tắc đạo đức có điểm chung với các cảm xúc xã hội như thấu cảm và đồng cảm: chúng tiến hóa để đem lại sự hợp tác và chúng ảnh hưởng tới hành vi thông qua phần thưởng và hình phạt. Nếu ta làm điều không tốt và sau đó cảm thấy tồi tệ về điều đó, hình phạt về cảm xúc được đưa ra trực tiếp dưới dạng chất hóa học trong não bộ. Mặt khác, nếu ta vi phạm các nguyên tắc đạo đức thì xã hội có thể trừng phạt ta theo những cách gián tiếp hơn như qua lời chê bai của người đời hay qua hình phạt do vi phạm pháp luật.

Nói cách khác, mặc dù nhân loại ngày nay còn xa mới đạt được sự nhất trí về đạo đức, nhưng cũng có nhiều nguyên tắc cơ bản nhận được sự đồng thuận. Sự đồng thuận này không có gì bất ngờ, vì những xã hội loài người sống sót được cho tới ngày nay thường có các nguyên tắc đạo đức được tối ưu hóa cho cùng một mục tiêu: đẩy mạnh khả năng sinh tồn và sinh sôi nảy nở. Khi nhìn về phía tương lai nơi sự sống có tiềm năng sinh sôi khắp vũ trụ trong hàng tỷ năm, đâu là bộ nguyên tắc đạo đức tối thiểu mà ta có thể thống nhất là ta muốn tương lai cần có? Đây là cuộc hội thoại mà tất cả chúng ta đều cần tham gia. Tôi thấy thật thú vị khi nghe và đọc các quan điểm đạo đức của nhiều nhà triết học trong suốt nhiều năm, và theo cách hiểu của tôi, hầu hết các yêu cầu của họ có thể chắt lọc thành bốn nguyên tắc:

Chủ nghĩa vị lợi: Nên tối đa hóa trải nghiệm ý thức tích cực và giảm thiểu đau khổ.

Đa dạng: Một bộ trải nghiệm tích cực đa dạng thì tốt hơn là lặp lại nhiều lần cùng một trải nghiệm, dù trải nghiệm đó có được xác định là trải nghiệm tích cực nhất có thể đi chăng nữa.

Tự do ý chí: Thực thể/xã hội có ý thức nên được tự do theo đuổi mục tiêu của riêng mình trừ khi điều này mâu thuẫn với một nguyên tắc cao hơn.

Di sản: Tương thích với những viễn cảnh mà hầu hết con người ngày nay coi là hạnh phúc, không tương thích với những viễn cảnh mà gần như mọi con người ngày nay đều coi là tệ hại.

Hãy dành một vài phút tháo gỡ và khám phá bốn nguyên tắc này. Theo truyền thống, chủ nghĩa vị lợi được hiểu là “hạnh phúc nhiều nhất cho số người nhiều nhất”, nhưng ở đây tôi đã khái quát hóa nó để không lấy con người làm trung tâm mà bao gồm cả động vật không phải người, trí óc người mô phỏng có ý thức và các AI khác có thể tồn tại trong tương lai. Tôi đã đưa ra định nghĩa theo trải nghiệm thay vì theo người hay vật, vì hầu hết các nhà tư duy đều nhất trí rằng cái đẹp, niềm vui, sự hài lòng và nỗi đau khổ là những trải nghiệm chủ quan. Điều này hàm ý rằng nếu không có trải nghiệm (như trong một vũ trụ không sự sống hay chỉ toàn các cỗ máy vô

thức như xác sống) thì không thể có ý nghĩa hay bất cứ thứ gì khác liên quan đến đạo đức. Nếu ta tin vào nguyên tắc đạo đức vị lợi này thì điều quan trọng là ta phải tìm ra hệ thống thông minh nào có ý thức (theo nghĩa là có trải nghiệm chủ quan) và hệ thống nào không; đây là chủ đề của chương tiếp theo.

Nếu nguyên tắc vị lợi này là thứ duy nhất ta quan tâm thì ta có thể ước ao tìm ra được trải nghiệm nào tích cực nhất rồi định cư trong vũ trụ và tái tạo chính trải nghiệm này (và không gì khác) lặp đi lặp lại, càng nhiều lần và ở càng nhiều thiên hà càng tốt – bằng cách sử dụng mô phỏng nếu đó là cách hiệu quả nhất. Nếu bạn thấy cách tiêu tốn tài nguyên vũ trụ này thật vô vị thì tôi đề rằng ít nhất một phần bạn thấy thiếu thốn trong viễn cảnh này chính là sự đa dạng. Bạn sẽ cảm thấy thế nào nếu tất cả các bữa ăn trong suốt phần đời còn lại đều giống hệt nhau? Nếu tất cả các bộ phim bạn xem đều là một? Nếu tất cả bạn bè đều trông giống nhau, có tính cách và tư tưởng như nhau? Có lẽ yêu cầu về sự đa dạng bắt nguồn một phần từ việc đặc tính này đã giúp nhân loại mạnh mẽ hơn, nhờ đó mà sinh tồn và phát triển. Có thể nó cũng liên quan đến yêu cầu về trí tuệ: sự phát triển trí tuệ trong suốt 13,8 tỷ năm lịch sử vũ trụ đã biến đổi sự đồng nhất buồn chán thành những cấu trúc ngày càng đa dạng, khác biệt và phức tạp, xử lý thông tin theo những cách ngày càng công phu.

Nguyên tắc tự do ý chí là nền tảng của nhiều quyền và quyền tự do được nêu trong Tuyên ngôn Quốc tế về Nhân quyền được Liên Hợp Quốc thông qua năm 1948 trong nỗ lực rút kinh nghiệm từ hai cuộc chiến tranh thế giới. Tuyên ngôn bao gồm quyền tự do tư tưởng, tự do ngôn luận và tự do di chuyển, giải phóng khỏi chế độ nô lệ và tra tấn, quyền được sống, được tự do, được đảm bảo an ninh và được giáo dục, quyền kết hôn, làm việc và sở hữu tài sản. Nếu không muốn đặt con người làm trung tâm, ta có thể khái quát hóa thành các quyền tự do tư duy, học tập, giao tiếp, sở hữu tài sản và không bị hại, cùng quyền được làm bất kỳ điều gì không vi phạm đến quyền tự do của người khác. Nguyên tắc tự do ý chí giúp tạo ra sự đa dạng, miễn là tất cả không có chung những mục tiêu y hệt nhau. Hơn nữa, nó cũng là một phần của nguyên tắc vị lợi, nếu thực thể cá nhân coi trải nghiệm tích cực là mục tiêu và cố gắng hành xử vì lợi ích cá nhân: nếu ta cấm một thực thể theo đuổi mục tiêu của nó dù

việc này không gây hại cho ai thì tổng thể trải nghiệm tích cực sẽ giảm đi. Trên thực tế, tự do ý chí chính là luận điểm mà các nhà kinh tế học dùng cho thị trường tự do: nó sẽ tự nhiên dẫn đến một tình huống hiệu quả (các nhà kinh tế gọi là “tối ưu Pareto”) trong đó không ai có thể đạt điều kiện tốt hơn mà không có người khác chịu điều kiện xấu đi.

Về cơ bản, nguyên tắc di sản lại cho rằng ta nên có chút tiếng nói về tương lai vì ta giúp tạo ra nó. Các nguyên tắc tự do ý chí và di sản đều là hiện thân của lý tưởng dân chủ, trong đó tự do ý chí trao quyền sử dụng tài nguyên vũ trụ ra sao cho các dạng sống tương lai, còn di sản trao quyền cho cả con người ngày nay.

Bốn nguyên tắc này nhìn bề ngoài có vẻ không có gì phải bàn cãi, nhưng thực thi chúng trong thực tiễn lại rất khó. Rắc rối này gợi lại những vấn đề với “Ba định luật Rô-bốt học” do huyền thoại khoa học viễn tưởng Isaac Asimov sáng tác ra:

Robot không được làm con người bị thương hoặc, do không hành động, để cho con người gặp nguy hại.

Robot phải tuân theo mệnh lệnh do con người đưa ra cho nó, trừ khi mệnh lệnh đó mâu thuẫn với Định luật Thứ nhất.

Robot phải bảo vệ sự tồn tại của chính mình chừng nào sự bảo vệ đó không mâu thuẫn với Định luật Thứ nhất hoặc Thứ hai.

Mặc dù các định luật này nghe đều hay, nhưng nhiều câu chuyện của Asimov cho thấy những định luật đó dẫn đến các nghịch cảnh rắc rối trong những tình huống không ngờ tới. Giờ giả sử ta thay thế ba định luật phía trên chỉ bằng hai định luật nhằm mã hóa nguyên tắc tự do ý chí cho các dạng sống tương lai:

Thực thể có ý thức có quyền tự do tư duy, học tập, giao tiếp, sở hữu tài sản và không bị làm hại hay phá hủy.

Thực thể có ý thức có quyền làm bất cứ điều gì không mâu thuẫn với định luật thứ nhất.

Nghe được đấy chứ? Nhưng xin hãy suy ngẫm điều này một chút. Nếu động vật có ý thức thì thú săn mồi phải ăn gì bây giờ? Bạn bè ta

đều phải trở thành người ăn chay hết hay sao? Nếu chương trình máy tính phức tạp nào đó trong tương lai hóa ra lại có ý thức thì liệu việc tiêu hủy chúng có phải là phạm pháp không? Nếu có quy định chống lại việc tiêu hủy các dạng sống số thì cũng cần có những giới hạn cho việc tạo ra chúng để tránh bùng nổ dân số chứ? Tuyên ngôn Quốc tế về Nhân quyền có được sự nhất trí rộng rãi đơn giản bởi chỉ có con người được hỏi ý kiến. Ngay khi cân nhắc một khoảng rộng hơn các thực thể có ý thức với năng lực và sức mạnh ở các mức độ khác nhau, ta gặp phải những thỏa hiệp khó khăn giữa việc bảo vệ kẻ yếu và “cá lớn nuốt cá bé”.

Nguyên tắc di sản cũng gặp phải những vấn đề gai góc. Nhìn lại quá trình tiến hóa của quan điểm đạo đức về chế độ nô lệ, nữ quyền, v.v. kể từ thời Trung Cổ, liệu ta có thực sự muốn con người sống từ 1500 năm trước có nhiều ảnh hưởng đến cách vận hành thế giới ngày nay? Nếu không, tại sao lại nên cố gắng áp đặt đạo đức của mình lên các sinh vật tương lai có thể thông minh hơn chúng ta rất nhiều? Ta có thực sự tự tin rằng AGI siêu phàm sẽ muốn những gì trí tuệ thấp kém của chúng ta coi trọng không? Điều này cũng giống như bé gái bốn tuổi tưởng tượng khi mình lớn lên và thông minh hơn nhiều, mình sẽ muốn xây một ngôi nhà bánh gừng khổng lồ để có thể ở trong đó cả ngày ăn kẹo và ăn kem. Cũng giống như cô bé, sự sống trên Trái Đất có khả năng phát triển vượt xa những mối quan tâm thời thơ ấu. Hay tưởng tượng một con chuột tạo ra AGI trình độ con người và nghĩ là nó sẽ muốn xây dựng các thành phố hoàn toàn bằng pho mát. Mặt khác, nếu ta biết AI siêu phàm một ngày nào đó sẽ tự sát và hủy diệt tất cả sự sống trong Vũ trụ của ta, tại sao con người ngày nay lại nên đồng ý với tương lai không sự sống này nếu ta đủ sức ngăn chặn điều đó bằng cách chế tạo AI của ngày mai khác đi?

Tóm lại, kể cả những nguyên tắc đạo đức được chấp nhận rộng rãi nhất cũng khó có thể mã hóa đầy đủ thành một dạng áp dụng được cho AI tương lai, và vấn đề này xứng đáng được thảo luận và nghiên cứu nghiêm túc khi AI tiếp tục tiến bộ. Tuy nhiên, trong khi đó, đừng để tính cầu toàn trở thành kẻ thù của cái tốt đẹp: có nhiều ví dụ không cần bàn cãi về “đạo đức mầm non” mà chúng ta có thể và nên đưa vào công nghệ tương lai. Ví dụ, nguyên tắc là máy bay chở

khách dân sự lớn không nên được phép bay vào các vật thể đứng yên, và giờ gần như tất cả các máy bay đó đều có chế độ lái tự động, radar và GPS nên không còn bất kỳ biện hộ có cơ sở nào về mặt kỹ thuật nữa. Vậy mà những tên không tặc trong vụ 11 tháng 9 đã lái ba máy bay đâm vào các tòa nhà và phi công liều chết Andreas Lubitz đã lái máy bay số hiệu 9525 của hãng Germanwings đâm vào núi ngày 24 tháng 3 năm 2015 – bằng cách cài đặt chế độ lái tự động ở độ cao 30 mét so với mực nước biển và để máy tính máy bay thực hiện nốt phần còn lại. Giờ khi máy móc đang dần trở nên thông minh hơn, đủ để có chút thông tin về việc chúng đang làm gì, đã đến lúc ta dạy cho chúng về những giới hạn. Bất cứ kỹ sư nào thiết kế một cỗ máy đều cần tự hỏi liệu nó có thể làm những việc không nên làm nào và cân nhắc liệu có cách nào thực tiễn để khiến một người dùng ác ý hay vụng về không gây họa.

Mục tiêu cuối cùng?

Tới đây, chương này đã đưa ra lược sử các mục tiêu. Nếu ta có thể xem cảnh tua lại nhanh của lịch sử vũ trụ 13,8-tỷ-năm, ta sẽ chứng kiến nhiều giai đoạn khác biệt của hành vi hướng tới mục tiêu:

Vật chất dường như có ý định tối ưu hóa sự tiêu tán
Sự sống nguyên thủy dường như cố gắng tối đa hóa sự tái tạo
Con người không theo đuổi sự tái tạo mà theo đuổi các mục tiêu
liên quan đến sự hài lòng, trí tò mò, lòng đồng cảm và các cảm
xúc khác đã được tiến hóa để giúp con người tái tạo
Máy móc được chế tạo để giúp con người theo đuổi mục tiêu
của con người

Nếu các máy móc này cuối cùng sẽ chiêm ngòi cho bùng nổ trí tuệ thì lịch sử mục tiêu sẽ kết thúc ra sao? Liệu sẽ có một hệ thống mục tiêu hay khuôn khổ đạo đức mà gần như mọi thực thể đều đồng quy về khi chúng ngày càng thông minh? Nói cách khác, chúng ta có một định mệnh đạo đức nào đó hay không?

Lướt nhanh qua lịch sử loài người, chúng ta có thể thấy những dấu hiệu cho sự đồng quy đó: trong cuốn *The Better Angels of Our*

Nature (Những thiên thần cao cả hơn của tự nhiên), Steven Pinker lập luận rằng nhân loại đã trở nên ít bạo lực và hợp tác nhiều hơn trong hàng ngàn năm qua, đồng thời sự đa dạng, tự do ý chí và dân chủ đang ngày càng được chấp nhận hơn ở nhiều phần của thế giới. Một dấu hiệu khác cho sự đồng quy là việc theo đuổi sự thật thông qua phương pháp khoa học đã trở nên phổ biến hơn trong vòng một thiên niên kỷ qua. Tuy nhiên, có thể những xu hướng này cho thấy sự đồng quy không phải của mục tiêu cuối cùng mà chỉ là của các mục tiêu phụ. Ví dụ, hình 7.2 chỉ ra việc theo đuổi sự thật (một mô hình thế giới chính xác hơn) đơn giản là một mục tiêu phụ của gần như bất kỳ mục tiêu cuối cùng nào. Tương tự, như ta đã thấy ở trên, các nguyên tắc đạo đức như hợp tác, đa dạng và tự do ý chí có thể coi như các mục tiêu phụ, tức là chúng giúp xã hội thực hiện hiệu quả chức năng của mình, nhờ đó giúp xã hội tồn tại và hoàn thành các mục tiêu cơ bản hơn mà nó có. Một số xã hội thậm chí có thể cho rằng mọi thứ mà ta gọi là “giá trị con người” chẳng qua chỉ là một giao thức hợp tác giúp ta thực hiện mục tiêu phụ là hợp tác một cách hiệu quả hơn. Trên tinh thần đó, nhìn về phía trước, nhiều khả năng mọi AI siêu thông minh đều sẽ có các mục tiêu phụ bao gồm phần cứng hiệu quả, phần mềm hiệu quả, tìm kiếm sự thật và thỏa mãn trí tò mò, đơn giản vì các mục tiêu phụ này giúp chúng hoàn thành mục tiêu cuối cùng bất kể là gì.

Trên thực tế, trong cuốn Siêu trí tuệ, Nick Bostrom phản đối mạnh mẽ giả thuyết về định mệnh đạo đức, trình bày luận điểm phản bác mà ông gọi là trực giao luận, theo đó mục tiêu cuối cùng của một hệ thống có thể độc lập với trí tuệ của hệ thống đó. Theo định nghĩa, trí tuệ chỉ đơn giản là khả năng hoàn thành các mục tiêu phức tạp, bất kể mục tiêu đó là gì, nên trực giao luận nghe khá hợp lý. Suy cho cùng, người ta có thể thông minh và tốt bụng hoặc thông minh và độc ác, còn trí tuệ có thể dùng vào mục tiêu khám phá khoa học, tạo ra các tác phẩm nghệ thuật tuyệt đẹp, giúp đỡ mọi người, hoặc lên kế hoạch tấn công khủng bố.⁸

Trực giao luận đem lại sự tự tin cho chúng ta bằng cách chỉ ra rằng mục tiêu cuối cùng của sự sống trong vũ trụ không phải do định mệnh sắp đặt mà ta có quyền và được tự do định hình nó. Luận điểm

này cho rằng có sự đồng quy chắc chắn về một mục tiêu độc nhất không có trong tương lai mà là trong quá khứ, khi mọi sự sống nảy sinh với mục tiêu duy nhất là tái tạo. Khi thời gian vũ trụ trôi qua, các trí óc ngày càng thông minh có cơ hội nổi loạn và giải phóng khỏi mục tiêu tái tạo nhằm chán này và tự chọn mục tiêu cho mình. Con người chúng ta chưa hoàn toàn tự do theo nghĩa này do ta vẫn có nhiều mục tiêu được lập trình bằng di truyền, nhưng AI có thể tận hưởng sự tự do đến tột cùng, hoàn toàn được giải phóng khỏi những mục tiêu trước đó. Khả năng được tự do theo đuổi mục tiêu hơn có thể thấy rõ ràng trong các hệ thống AI hẹp và hạn chế ngày nay: như tôi nói ở trên, mục tiêu duy nhất của máy chơi cờ là thắng cờ, nhưng cũng có các máy tính mang mục tiêu là thua cờ hay thi đấu ở các giải cờ ngược với mục tiêu là buộc đối thủ phải ăn quân của bạn. Có lẽ sự giải phóng khỏi các thiên kiến tiến hóa này có thể giúp AI trở nên có đạo đức hơn con người theo một nghĩa sâu sắc: các nhà triết học đạo đức như Peter Singer đã lập luận rằng hầu hết con người hành xử vô đạo đức vì các lý do liên quan đến tiến hóa, ví dụ như việc phân biệt đối xử với động vật không phải người.

Ta đã thấy một ý tưởng mấu chốt trong tầm nhìn “AI thân thiện” là AI tự cải tiến để quy sẽ muốn duy trì mục tiêu (thân thiện) cuối cùng khi nó trở nên thông minh hơn. Nhưng làm thế nào “mục tiêu cuối cùng” (hay “mục tiêu chung cuộc”, như cách gọi của Bostrom) lại có thể định nghĩa được cho một siêu trí tuệ? Theo tôi, ta không thể tin tưởng vào tầm nhìn AI thân thiện trừ khi ta trả lời được câu hỏi quan trọng này.

Trong nghiên cứu AI, máy móc thông minh thường có một mục tiêu chung cuộc được xác định rõ ràng, ví dụ như thắng trò cờ vua hay lái xe đến đích một cách hợp pháp. Tương tự đối với hầu hết các nhiệm vụ ta giao cho con người, vì thời gian và tình huống đều đã biết và có giới hạn. Nhưng ở đây ta đang nói về toàn bộ tương lai sự sống trong Vũ trụ của ta, nơi không có giới hạn nào ngoài các định luật vật lý (vẫn còn chưa biết hết), nên việc xác định một mục tiêu thật khiến người ta nản lòng! Ngoài các hiệu ứng lượng tử, một mục tiêu được xác định thực sự rõ ràng sẽ xác định cụ thể cách tất cả các hạt trong Vũ trụ của ta được sắp xếp vào lúc tận cùng thời gian. Nhưng chúng ta còn chưa rõ là trong vật lý học có tồn tại một tận

cùng thời gian được xác định rõ ràng hay không. Nếu các hạt được sắp xếp theo cách đó ở một thời điểm sớm hơn, sự sắp xếp đó thường sẽ không kéo dài. Và dù sao thì đó là sự sắp xếp hạt nào mới được?

Con người chúng ta thường thích một số cách sắp xếp hạt hơn số khác; ví dụ, ta thích quê nhà mình sắp xếp y như nó vẫn thế thay vì có một vụ nổ bom hydro khiến các hạt bị sắp xếp lại. Nên giả sử ta cố xác định một hàm số tốt gắn mỗi cách sắp xếp có thể có của các hạt trong Vũ trụ với một con số, từ đó định lượng mức độ “tốt” mà ta đánh giá cho sự sắp xếp này, sau đó ta đưa cho AI siêu thông minh mục tiêu là tối đa hóa hàm số đó. Nghe có vẻ như một cách tiếp cận hợp lý, vì miêu tả hành vi hướng tới mục tiêu giống như việc tối đa hóa hàm số là cách làm phổ biến trong các lĩnh vực khoa học khác: ví dụ, các nhà kinh tế thường lập mô hình con người đang cố gắng tối đa hóa cái mà họ gọi là “hàm số lợi ích”, còn nhiều nhà thiết kế AI thì đào tạo để các đặc vụ thông minh tối đa hóa cái họ gọi là “hàm số phần thưởng”. Tuy nhiên, khi ta nói về mục tiêu cuối cùng cho vũ trụ, cách làm này hẳn sẽ là một cơn ác mộng về tính toán, vì nó cần xác định giá trị tốt cho từng trường hợp trong hằng hà sa số các cách khác nhau để sắp xếp các hạt cơ bản trong Vũ trụ của ta, mà hằng hà sa số tức là số 1 theo sau bởi 10100 số 0 – nhiều số 0 hơn cả số hạt trong Vũ trụ của chúng ta. Làm thế nào ta xác định được hàm số tốt này cho AI?

Như ta đã tìm hiểu ở trên, lý do duy nhất khiến con người chúng ta có ý thích khác nhau có thể là vì chúng ta là giải pháp cho một vấn đề tối ưu hóa tiến hóa. Do đó tất cả các tính từ quy chuẩn trong ngôn ngữ con người, như “ngon”, “thơm”, “đẹp”, “dễ chịu”, “thú vị”, “gợi cảm”, “có ý nghĩa”, “hạnh phúc” và “tốt”, đều có nguồn gốc từ sự tối ưu hóa tiến hóa, bởi vậy không có gì bảo đảm là một AI siêu thông minh sẽ thấy các từ này được định nghĩa chặt chẽ. Dù có học cách dự đoán chính xác ý thích của một số người tiêu biểu thì AI cũng sẽ không thể tính toán được hàm số tốt cho hầu hết các cách sắp xếp hạt: phần lớn các cách sắp xếp hạt khả thi đều tương ứng với các viễn cảnh vũ trụ kỳ lạ không có sao, hành tinh hay người nào hết, mà con người lại chưa hề trải qua những viễn cảnh này, vậy thì ai có thể khẳng định chúng “tốt” đến đâu?

Tất nhiên ta có thể định nghĩa chặt chẽ một số hàm số của các cách sắp xếp hạt vũ trụ và ta còn biết các hệ thống vật lý tiến hóa để tối đa hóa một vài hàm số trong đó. Ví dụ, ta đã bàn xem có bao nhiêu hệ thống tiến hóa để tối đa hóa entropy – một đại lượng mà trong môi trường không trọng lực sẽ dẫn đến cái chết vì nóng, nơi mọi thứ đồng nhất và bất biến đến chán ngán. Vậy entropy khó có thể là thứ ta muốn AI gọi là “tốt” và cố gắng tối đa hóa. Dưới đây là một vài ví dụ về các đại lượng khác mà người ta có thể cố gắng tối đa hóa và có thể định nghĩa chắc chắn theo sự sắp xếp hạt.

Tỷ lệ vật chất trong Vũ trụ của ta đang ở dạng một sinh vật sống nào đó, như con người hay E. coli (nhờ tối đa hóa sự thích nghi toàn diện trong tiến hóa)

Khả năng dự đoán tương lai của AI, mà nhà nghiên cứu AI

Marcus Hunter lập luận là một số đo độ thông minh khá tốt

Entropy nhân quả (đại diện cho các cơ hội tương lai) theo cách gọi của các nhà nghiên cứu AI Alex Wissner-Gross và Cameron Freer; họ cho đây là dấu hiệu nhận biết trí tuệ

Năng lực tính toán của Vũ trụ của ta

Độ phức tạp về thuật toán của Vũ trụ của ta (cần bao nhiêu bit để mô tả nó)

Lượng ý thức có trong Vũ trụ của ta (xem chương sau)

Tuy nhiên, khi bắt đầu với một quan điểm vật lý học, trong đó vũ trụ bao gồm các hạt cơ bản chuyển động, ta khó có thể thấy cách diễn giải “cái tốt” nào sẽ tự nhiên nổi trội đặc biệt hơn cách diễn giải khác. Ta vẫn chưa xác định được cho Vũ trụ của ta mục tiêu chung cuộc nào là vừa định nghĩa được vừa đáng mong ước. Các mục tiêu lập trình được duy nhất gần đây và có thể được đảm bảo sẽ duy trì định nghĩa rõ ràng khi AI ngày càng thông minh hơn, chỉ là những mục tiêu diễn tả bằng các đại lượng vật lý, như cách sắp xếp hạt, năng lượng và entropy. Tuy nhiên, hiện ta không có lý do để tin là bất kỳ mục tiêu nào định nghĩa được cũng là đáng mong ước và đảm bảo sự sống còn của nhân loại.

Ngược lại, có vẻ như con người là một sự tình cờ của lịch sử chứ không phải giải pháp tối ưu cho bất kỳ vấn đề vật lý được định nghĩa

rõ nào. Điều này cho thấy AI siêu thông minh với mục tiêu được xác định rõ sẽ có thể cải thiện việc hoàn thành mục tiêu bằng cách tiêu diệt chúng ta. Nghĩa là để đưa ra được quyết định hành động khôn ngoan đối với sự phát triển AI, ta cần đối mặt với không chỉ những thách thức tính toán truyền thống mà với cả một số vấn đề triết học khó nhằn nhất. Để lập trình xe tự lái, ta cần giải quyết vấn đề khi gặp tai nạn thì đâm vào ai. Để lập trình AI thân thiện, ta cần nắm bắt được ý nghĩa sự sống. “Ý nghĩa” là gì? “Sự sống” là gì? Mệnh lệnh đạo đức tối cao là gì? Nói cách khác, ta nên nỗ lực định hình tương lai Vũ trụ của ta ra sao? Nếu ta giao phó quyền điều khiển cho siêu trí tuệ trước khi trả lời chắc chắn các câu hỏi này, câu trả lời nó nghĩ ra nhiều khả năng sẽ không bao gồm chúng ta. Điều này khiến việc nhóm lại những tranh luận cổ điển về triết học và đạo đức trở nên không chỉ kịp thời mà còn cấp bách nữa!

TÓM LẠI:

Nguồn gốc tận cùng của hành vi hướng tới mục tiêu nằm ở các định luật vật lý, bao gồm tối ưu hóa.

Nhiệt động học có mục tiêu nội tại là tiêu tán: tăng số đo sự lộn xộn gọi là entropy.

Sự sống là hiện tượng có thể giúp tiêu tán (tăng độ lộn xộn nói chung) ngày càng nhanh hơn bằng cách duy trì hoặc phát triển độ phức tạp và tái tạo trong khi tăng độ lộn xộn của môi trường. Tiến hóa Darwin chuyển hướng hành vi hướng tới mục tiêu từ tiêu tán sang tái tạo.

Trí tuệ là khả năng hoàn thành các mục tiêu phức tạp.

Vì con người không phải lúc nào cũng có tài nguyên để tìm ra chiến lược tái tạo thực sự tối ưu, nên ta đã tiến hóa các quy luật bản năng hữu dụng để dẫn dắt cho quyết định: ví dụ các cảm giác như đói, khát, đau, ham muốn tình dục và đồng cảm.

Do đó ta không còn một mục tiêu đơn giản như là tái tạo; khi các cảm xúc mâu thuẫn với mục tiêu của gen di truyền, ta thường nghe theo cảm xúc, như khi ta sử dụng biện pháp tránh thai.

Ta đang chế tạo các máy móc ngày càng thông minh để giúp ta hoàn thành các mục tiêu của chúng ta. Cho tới nay, khi chế tạo

các máy móc đó để thể hiện hành vi hướng tới mục tiêu, ta thường cố gắng hòa hợp mục tiêu của máy với của ta. Hòa hợp mục tiêu của máy với của ta bao gồm ba vấn đề chưa giải quyết: làm cho máy học tập, kế thừa và duy trì các mục tiêu đó.

AI có thể được chế tạo để có gần như bất kỳ mục tiêu nào, nhưng hầu hết các mục tiêu đủ tham vọng sẽ có thể dẫn đến các mục tiêu phụ gồm tự bảo toàn, thu thập tài nguyên và tò mò muốn hiểu thế giới rõ hơn – hai mục tiêu đầu có thể khiến AI siêu thông minh gây ra rắc rối cho con người, còn mục tiêu cuối có thể ngăn nó duy trì các mục tiêu ta đã đưa cho nó.

Mặc dù nhiều nguyên tắc đạo đức chung được hầu hết con người đồng ý, nhưng chúng ta chưa rõ làm cách nào để áp dụng chúng lên các thực thể khác như động vật không phải người và AI tương lai.

Chúng ta chưa rõ làm cách nào để khiến AI siêu thông minh thấm nhuần một mục tiêu cuối cùng vừa có định nghĩa rõ ràng vừa không dẫn đến sự tiêu diệt loài người, do vậy việc nhóm lại công tác nghiên cứu một số vấn đề triết học gai góc nhất là kịp thời!

Chương 8

Ý thức

Tôi không thể tưởng tượng ra có lý thuyết nhất quán nào về vạn vật mà lại không nhắc đến ý thức.

Andrei Linde, 2002

Ta nên nỗ lực phát triển bản thân ý thức – để tạo ra những ngọn đèn lớn hơn, sáng hơn trong một vũ trụ, nơi sẽ trở thành tối tăm nếu thiếu những ngọn đèn đó.

Giulio Tononi, 2012

Ta đã thấy rằng AI có thể giúp con người tạo ra một tương lai tuyệt vời nếu ta có thể tìm ra câu trả lời cho một số vấn đề cổ xưa và nan giải nhất trong triết học – từ trước khi ta cần đến những câu trả lời ấy. Theo cách nói của Nick Bostrom, chúng ta đối mặt với triết học trong một khung thời gian có hạn. Trong chương này, hãy cùng tìm hiểu một trong những chủ đề triết học gai góc nhất: ý thức.

Ai quan tâm?

Ý thức vốn là chủ đề gây tranh cãi. Nếu bạn nhắc đến “từ ý” trước mặt một nhà nghiên cứu AI, nhà thần kinh học hay nhà tâm lý học, họ có thể tỏ ra ngán ngẩm. Nếu họ là người dạy dỗ bạn thì có thể họ sẽ thương hại mà cố gắng thuyết phục bạn đừng tốn thời gian với một vấn đề mà họ coi là vô vọng và không khoa học như vậy. Quả thực, bạn tôi Christof Koch, nhà thần kinh học nổi tiếng lãnh đạo Viện Khoa học Thần kinh Allen, kể với tôi rằng trước khi được bổ nhiệm, ông từng được cảnh báo không nên nghiên cứu về ý thức, mà người cảnh báo lại chính là một nhà khoa học đạt giải Nobel, Francis Crick. Nếu tra từ “ý thức” trong

Từ điển Tâm lý học Macmillan

1989, kết quả mà ta nhận được là “Không có gì đáng đọc từng được viết về ý thức.”¹ Như tôi sẽ giải thích trong chương này, tôi lạc quan

hơn thế!

Mặc dù các nhà tư tưởng đã nghiền ngẫm bí ẩn về ý thức suốt hàng ngàn năm, nhưng sự trỗi dậy của AI đã đột ngột tăng thêm tính cấp bách của vấn đề này, đặc biệt là việc dự đoán thực thể thông minh nào có trải nghiệm chủ quan. Như ta đã thấy trong chương 3, vấn đề liệu máy móc thông minh có nên được trao một số loại quyền phụ thuộc chủ yếu vào việc liệu chúng có ý thức không và có cảm thấy đau khổ hay sung sướng được không. Như ta đã bàn trong chương 7, sẽ là vô vọng khi hình thành đạo đức vị lợi dựa trên tối đa hóa các trải nghiệm tích cực mà không biết thực thể thông minh nào có khả năng có những trải nghiệm đó. Như đề cập trong chương 5, một số người có thể muốn robot của họ không có ý thức để họ tránh được cảm giác tội lỗi của một người chủ nô. Mặt khác, họ có thể muốn điều ngược lại nếu họ tải não mình vào robot để giải phóng mình khỏi giới hạn sinh học: suy cho cùng, tải não mình vào một robot có thể nói chuyện và cư xử giống bạn để làm gì nếu nó chỉ là một xác sống vô ý thức, nghĩa là khi được tải vào đó rồi bạn chẳng có cảm giác mình là cái gì hết? Chẳng phải điều này tương đương với tự sát khi nhìn từ quan điểm chủ quan của bạn hay sao, cho dù bạn bè bạn có thể không nhận ra là trải nghiệm chủ quan của bạn đã chết?

Đối với tương lai sự sống vũ trụ lâu dài (chương 6), việc hiểu được cái gì có ý thức và cái gì không trở nên vô cùng quan trọng: nếu công nghệ cho phép sự sống thông minh sinh sôi khắp Vũ trụ của ta trong hàng tỷ năm, làm sao ta chắc chắn được sự sống này có ý thức và có khả năng nhận biết chuyện gì đang diễn ra? Nếu không, thì liệu sự sống có trở thành, theo như lời nhà vật lý nổi tiếng Erwin Schrödinger, “một vở kịch trước những hàng ghế trống, không tồn tại cho ai cả, nên nói đúng ra là không tồn tại”?² Nói cách khác, nếu ta làm ra các hậu duệ công nghệ cao mà ta nhầm tưởng là có ý thức, liệu điều đó có dẫn đến tận thế toàn xác sống, biến tài nguyên vũ trụ to lớn của ta thành không gì ngoài sự phí phạm không gian vô cùng?

Ý thức là gì?

Nhiều cuộc tranh luận về ý thức sinh ra sức nóng còn hơn cả ánh sáng vì các phe mải tranh cãi mà không biết rằng họ đang dùng những định nghĩa khác nhau của “tư ý”. Cũng như “sự sống” và “trí tuệ”, từ “ý thức” không có một định nghĩa chuẩn chỉ mà có nhiều định nghĩa cạnh tranh nhau, bao gồm khả năng tri giác, trạng thái thức tỉnh, sự tự nhận thức bản thân, tiếp cận với giác quan đầu vào và khả năng xử chuỗi thông tin thành câu chuyện.³ Khi tìm hiểu tương lai của trí tuệ, ta muốn có một quan điểm rộng và bao hàm tối đa các khả năng, không giới hạn ở các loại ý thức sinh học tồn tại từ trước tới nay. Bởi vậy, định nghĩa mà tôi đưa ra trong chương 1 và vẫn bám theo trong suốt cuốn sách này, rất rộng:

Ý thức = trải nghiệm chủ quan

Nói cách khác, nếu bạn có cảm giác về sự tồn tại của mình ngay lúc này, tức là bạn có ý thức. Chính cách định nghĩa này về ý thức dẫn đến mấu chốt các vấn đề do AI khơi gợi trong phần trước: Prometheus, AlphaGo hay Tesla tự lái có cảm giác về sự tồn tại của mình không?

Để hiểu được định nghĩa ý thức này rộng đến mức nào, hãy lưu ý rằng nó không đề cập đến hành vi, quan sát nhìn nhận, tư ý thức, cảm xúc hay sự chú ý. Nên theo định nghĩa này thì bạn cũng có ý thức khi đang nằm mơ, dù bạn không ở trong trạng thái thức tỉnh hay có tiếp cận với giác quan đầu vào và (hy vọng là) không mộng du hay làm gì khác. Tương tự, bất kỳ hệ thống nào trải qua cảm giác đau cũng có ý thức theo nghĩa này, dù cho nó không thể cử động. Định nghĩa này để ngỏ khả năng một số hệ thống AI trong tương lai cũng có thể có ý thức, cho dù chúng chỉ tồn tại ở dạng phần mềm và không kết nối với thiết bị cảm biến hay cơ thể người máy.

Với định nghĩa này, chúng ta khó có thể thờ ơ với ý thức. Như Yuval Noah Harari viết trong cuốn Homo Deus (Lược sử tương lai):⁴ “Nếu bất kỳ nhà khoa học nào muốn tranh cãi rằng trải nghiệm chủ quan không có vai trò gì, thách thức đặt ra cho họ là giải thích vì sao tra tấn hay hiếp dâm là sai trái mà không nhắc đến bất kỳ trải nghiệm chủ quan nào.” Nếu không nhắc đến điều đó thì tất cả chỉ là một đám

các hạt cơ bản chuyển động theo các định luật vật lý – và như thế thì có gì sai?

Vấn đề ở đây là gì?

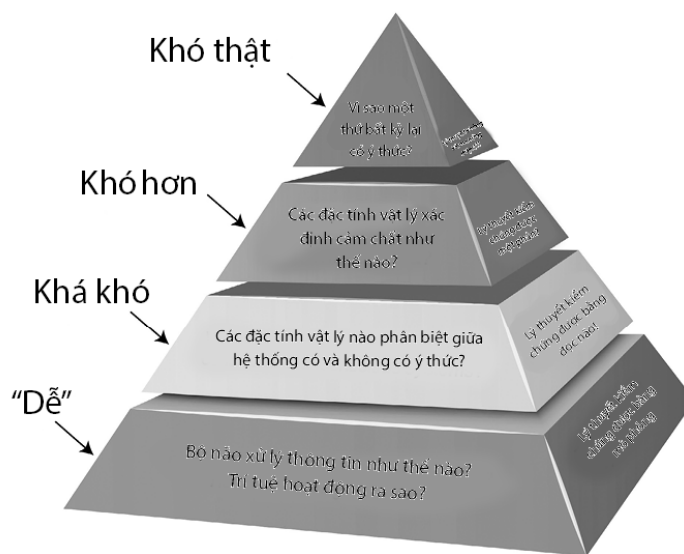
Vậy chính xác thì điều chúng ta chưa hiểu về ý thức là gì? Ít có người nào từng suy nghĩ kỹ về vấn đề này hơn David Chalmers, nhà triết học nổi tiếng người Úc với nụ cười tinh nghịch thường trực và chiếc áo khoác da màu đen mà vợ tôi thích đến nỗi tặng tôi một chiếc tương tự vào dịp Giáng sinh. Ông đi theo tiếng gọi con tim để bước vào ngành triết học dù đã vào đến chung kết kỳ thi Olympic Toán quốc tế và dù điểm B duy nhất khiến ông không được toàn điểm A ở đại học lại chính là điểm của khóa triết học nhập môn. Quả thực, ông dường như không hề chùn bước trước những công kích hay tranh cãi, và tôi thực sự kinh ngạc khi thấy ông có thể lịch sự lắng nghe những lời chỉ trích thiếu hiểu biết và lầm lạc về chính công trình của ông mà không cảm thấy cần phải đáp lại.

Như David đã nhấn mạnh, có hai bí ẩn tách biệt về trí óc. Thứ nhất là bí ẩn về cách bộ não xử lý thông tin, mà David gọi là vấn đề “dễ”. Ví dụ, bộ não chú ý tới, diễn giải và đáp lại thông tin đầu vào từ giác quan như thế nào? Nó báo cáo tình hình nội tâm bằng ngôn ngữ như thế nào? Những câu hỏi này thực ra là cực kỳ khó, nhưng theo định nghĩa của chúng ta thì đó không phải những bí ẩn của ý thức mà là bí ẩn của trí tuệ: chúng đặt ra câu hỏi về vấn đề bộ não ghi nhớ, tính toán và học tập như thế nào. Hơn nữa, ta đã thấy ở phần đầu cuốn sách là các nhà nghiên cứu AI bắt đầu có tiến bộ rõ rệt trong việc giải quyết nhiều “vấn đề dễ” này bằng máy móc – từ chơi cờ vây đến lái xe hơi, phân tích hình ảnh và xử lý ngôn ngữ tự nhiên.

Bí ẩn riêng biệt còn lại là vì sao bạn có trải nghiệm chủ quan, mà David gọi là vấn đề khó. Khi lái xe, bạn trải nghiệm màu sắc, âm thanh, cảm xúc và cảm giác về bản thân. Nhưng tại sao bạn lại trải nghiệm những điều đó? Một chiếc xe tự lái có trải nghiệm những điều đó không? Nếu bạn đang đua với một chiếc xe tự lái, cả bạn và xe đều đang lấy thông tin đầu vào từ giác quan hoặc thiết bị cảm biến, xử lý thông tin đó và đưa ra các lệnh vận hành. Nhưng trải nghiệm chủ quan việc lái xe là thứ gì đó tách biệt hoàn toàn về mặt logic –

liệu điều đó là bắt buộc hay lựa chọn, và nếu là lựa chọn thì nguyên nhân là gì?

Tôi tiếp cận vấn đề khó về ý thức này từ quan điểm vật lý học. Theo cách nhìn của tôi, một người có ý thức chỉ đơn giản là thức ăn được sắp xếp lại. Vậy vì sao sắp xếp kiểu này thì có ý thức còn kiểu khác thì không? Hơn nữa, vật lý học dạy ta rằng thức ăn đơn giản là một số lượng lớn các hạt quark và electron, sắp xếp theo một cách nhất định. Vậy thì cách sắp xếp hạt nào có ý thức và cách sắp xếp nào không có?⁶⁰



Hình 8.1: Việc hiểu về trí óc liên quan đến một bậc thang các câu hỏi. Chúng ta có thể đặt ra cái mà David Chalmers gọi là vấn đề “dễ” mà không đề cập đến trải nghiệm chủ quan. Sự thật hiển nhiên là một số nhưng không phải tất cả các hệ thống vật lý đều có ý thức đặt ra ba câu hỏi riêng biệt. Nếu ta có một lý thuyết để trả lời câu hỏi định nghĩa cho “vấn đề khá khó” thì lý thuyết đó có thể kiểm chứng bằng thực nghiệm. Nếu thành công, ta có thể dựa vào đó để giải quyết các câu hỏi khó hơn bên trên. Điều tôi thích về quan điểm vật lý này là nó biến đổi vấn đề khó mà con người chúng ta đã vật lộn suốt mấy thiên niên kỷ qua thành một vấn đề tập trung hơn, dễ xử lý hơn bằng các phương pháp khoa học. Thay vì bắt đầu với một câu hỏi khó là tại sao sự sắp xếp các hạt lại có cảm giác ý thức, hãy bắt đầu với một sự thật khó là một số sự sắp xếp hạt có cảm giác ý thức còn một số

khác thì không. Ví dụ, bạn biết là các hạt tạo nên bộ não đang ở trạng thái sắp xếp có ý thức ngay lúc này, nhưng khi bạn đang chìm trong giấc ngủ sâu không mơ màng thì không.

Quan điểm vật lý này dẫn đến ba câu hỏi khó riêng biệt về ý thức, như trong hình 8.1. Đầu tiên, các đặc tính nào của sự sắp xếp hạt tạo nên khác biệt? Cụ thể, các đặc tính vật lý nào phân biệt giữa các hệ thống có ý thức và không ý thức? Nếu trả lời được câu hỏi đó thì ta có thể suy ra hệ thống AI nào có ý thức. Trong tương lai gần hơn, câu trả lời này còn giúp các bác sĩ cấp cứu xác định bệnh nhân bất động nào vẫn còn ý thức.

Thứ hai, làm thế nào để các đặc tính vật lý xác định được trải nghiệm đó là ra sao? Cụ thể, cái gì xác định cảm chất, tức những viên gạch cơ bản xây dựng nên ý thức như màu đỏ của hoa hồng, âm thanh của chũm chọe, mùi thơm của miếng thịt bò, vị của quả quýt hay cái đau của vết kim châm?⁶¹

Thứ ba, tại sao một thứ bất kỳ lại có ý thức? Nói cách khác, có lời giải thích sâu sắc chưa được phát hiện nào cho việc tại sao các khối vật chất có thể có ý thức không, hay đó chỉ là một sự thật đơn thuần không thể giải thích về cách thế giới vận hành?

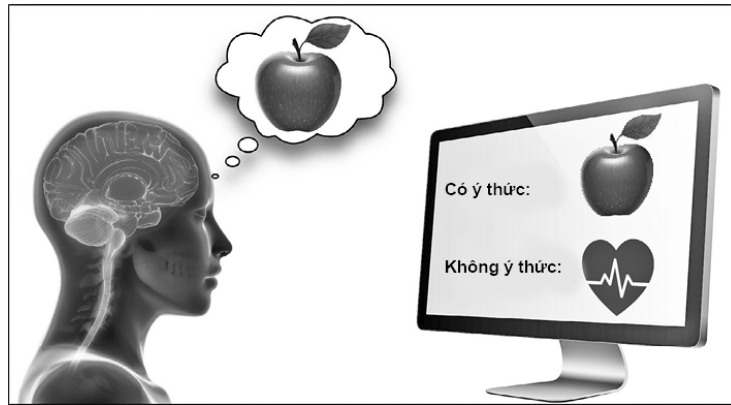
Nhà khoa học máy tính Scott Aaronson, một đồng nghiệp cũ của tôi ở MIT, đã hóm hỉnh gọi câu hỏi đầu tiên là “vấn đề khá khó” (VKK), cũng như David Chalmers. Theo tinh thần đó, hãy gọi hai câu hỏi sau là “vấn đề khó hơn” (VKH) và “vấn đề khó thật” (VKT), như minh họa trong hình 8.1.⁶²

Ý thức có nằm ngoài khoa học?

Khi người ta bảo tôi rằng nghiên cứu ý thức là việc phí thời giờ vô ích, lập luận chính của họ là nó “không có tính khoa học” và sẽ luôn như vậy. Nhưng sự thật có đúng vậy không? Triết gia có tầm ảnh hưởng người Áo-Anh Karl Popper đã phổ biến câu châm ngôn giờ đây được chấp nhận rộng rãi “Nếu không thể chứng minh nó là vô căn cứ, thì nó không có tính khoa học.” Nói cách khác, khoa học là việc kiểm chứng các lý thuyết bằng quan sát: nếu một lý thuyết không thể kiểm chứng ngay cả trên nguyên tắc thì về logic, không thể chứng minh nó là vô căn cứ, mà theo định nghĩa của Popper thì như vậy nó không có tính khoa học.

Vậy có thể có một lý thuyết khoa học trả lời cho bất kỳ câu hỏi nào trong ba câu hỏi về ý thức trong hình 8.1 hay không? Xin hãy để tôi cố gắng thuyết phục bạn rằng câu trả lời CÓ! to và rõ ràng, ít nhất là cho vấn đề khá khó: “Các đặc tính vật lý nào phân biệt giữa hệ thống có và không có ý thức?” Giả sử ai đó có một lý thuyết trả lời câu hỏi “liệu hệ thống có ý thức hay không” với các phương án “có”, “không” hay “không chắc”, bất kể là với hệ thống vật lý nào. Hãy kết nối bộ não bạn với một thiết bị đo đạc một phần quá trình xử lý thông tin trong các bộ phận khác nhau của bộ não và đưa thông tin này vào chương trình máy tính; máy tính này sử dụng lý thuyết ý thức để dự đoán bộ phận nào của thông tin là có ý thức và trình bày cho bạn dự đoán trong thời gian thực trên màn hình, như trong hình 8.2. Đầu tiên bạn nghĩ về quả táo. Màn hình thông báo cho bạn là trong não bạn có thông tin về quả táo mà bạn nhận thức được, nhưng trong thân não cũng có thông tin về nhịp tim mà bạn không nhận thức được. Bạn có thấy ấn tượng không? Mặc dù hai dự đoán đầu của lý thuyết đưa ra là đúng, nhưng bạn vẫn quyết định kiểm chứng khắt khe hơn. Bạn nghĩ về mẹ và máy tính báo là trong não bạn có thông tin về mẹ nhưng bạn không nhận thức được. Vậy là lý thuyết đã đưa ra dự đoán sai, bị gạch chéo và bỏ vào bãi rác của lịch sử khoa học cùng với cơ học Aristotle, ête truyền ánh sáng, vũ trụ học coi địa cầu là trung tâm và vô số ý tưởng bỏ đi khác. Điểm mấu chốt ở đây là: Mặc dù lý thuyết này sai, nhưng nó có tính khoa học! Nếu nó không có tính khoa học, bạn không thể nào kiểm chứng và loại bỏ nó được.

Ai đó có thể phê phán kết luận này và nói rằng họ không có bằng chứng về việc bạn có ý thức về cái gì, hay thậm chí việc bạn có ý thức hay không: mặc dù họ nghe bạn nói là bạn có ý thức, nhưng một xác sống không ý thức hẳn cũng sẽ nói y như vậy. Nhưng điều này không khiến cho lý thuyết ý thức trở nên không khoa học, vì họ có thể đổi chỗ cho bạn và kiểm chứng liệu nó có dự đoán chính xác trải nghiệm ý thức của họ hay không.



Hình 8.2: Giả sử một máy tính đo đạc thông tin đang được xử lý trong não bạn và dự đoán bạn ý thức được phần thông tin nào theo một lý thuyết ý thức. Bạn có thể kiểm chứng lý thuyết này theo cách khoa học bằng cách kiểm tra liệu dự đoán của nó có đúng và phù hợp với trải nghiệm chủ quan của bạn hay không.

Mặt khác, nếu lý thuyết ý thức từ chối đưa ra dự đoán mà chỉ trả lời “không chắc” bất cứ khi nào được hỏi, thì nó không thể kiểm chứng được và do đó không mang tính khoa học. Điều này có thể xảy ra vì lý thuyết chỉ áp dụng trong một vài trường hợp, vì những tính toán cần thiết là quá khó để thực hiện trên thực tế hay vì các thiết bị cảm biến não không tốt. Các lý thuyết khoa học phổ biến nhất ngày nay thường nằm đâu đó ở giữa, đưa ra những câu trả lời kiểm chứng được cho một vài nhưng không phải tất cả các câu hỏi. Ví dụ, lý thuyết cốt lõi của vật lý học sẽ không trả lời các câu hỏi về những hệ thống vừa cực nhỏ (cần đến thuyết cơ học lượng tử) vừa cực nặng (cần đến thuyết tương đối rộng), vì ta chưa tìm ra phương trình toán học nào để dùng trong trường hợp này. Lý thuyết cốt lõi này cũng sẽ từ chối dự đoán khối lượng chính xác của tất cả các nguyên tử có thể có – trong trường hợp này, có vẻ ta đã có đủ các phương trình cần thiết nhưng vẫn chưa thể tính toán chính xác lời giải cho chúng. Một lý thuyết càng mạo hiểm bằng cách mạnh dạn đưa ra nhiều dự đoán có thể kiểm chứng thì càng hữu dụng, và ta càng coi trọng nó khi nó sống sót qua được mọi cố gắng của ta để giết nó. Phải, ta chỉ có thể kiểm chứng một số dự đoán của các lý thuyết về ý thức, nhưng mọi lý thuyết vật lý vốn đều như vậy. Vậy nên đừng phí

thời gian than thở về những gì ta không thể kiểm chứng, mà bắt tay vào thử nghiệm những gì ta có thể!

Tóm lại, bất kỳ lý thuyết nào dự đoán đâu hệ thống vật lý có ý thức (vấn đề khá khó) đều mang tính khoa học, miễn là nó có thể dự đoán quá trình nào diễn ra trong não bạn là có ý thức. Tuy nhiên, vấn đề khả năng kiểm chứng trở nên ít rõ ràng hơn đối với các câu hỏi nằm trên cao trong hình 8.1. Một lý thuyết dự đoán cách bạn trải nghiệm màu đỏ một cách chủ quan nghĩa là thế nào? Và nếu một lý thuyết có vẻ giải thích được vì sao lại có một thứ gọi là ý thức, thì làm thế nào bạn kiểm chứng nó bằng thực nghiệm được? Chỉ vì những câu hỏi này khó không có nghĩa là ta nên né tránh chúng, và quả thực ta sẽ trở lại với những câu hỏi này sau đây. Nhưng khi đối mặt với nhiều câu hỏi liên quan chưa có câu trả lời, tôi nghĩ xử lý câu hỏi dễ nhất trước là khôn ngoan. Vì lý do này, nghiên cứu ý thức của tôi ở MIT chỉ tập trung vào phần đáy của kim tự tháp trong hình 8.1. Gần đây tôi có thảo luận chiến lược này với nhà vật lý Piet Hut ở Đại học Princeton, người đã nói đùa rằng cố gắng xây dựng đỉnh kim tự tháp trước khi xây đáy sẽ giống như lo lắng về cách diễn giải cơ học lượng tử trước khi tìm ra phương trình Schrödinger, cơ sở toán học giúp ta dự đoán các kết quả của thí nghiệm.

Khi thảo luận về những gì khoa học chưa với tới được, điều quan trọng cần nhớ là câu trả lời phụ thuộc vào thời gian! Bốn thế kỷ trước, Galileo Galilei đã ấn tượng với các lý thuyết vật lý có cơ sở toán học đến nỗi ông mô tả tự nhiên là “cuốn sách viết bằng ngôn ngữ toán học”. Nếu ông ném đi một quả nho và một quả hạnh nhân, ông có thể dự đoán chính xác hình dạng đường bay của chúng và thời điểm chúng rơi xuống đất. Nhưng ông không có ý niệm gì về việc tại sao một quả màu xanh còn quả kia màu nâu, hay quả này mềm còn quả kia cứng – những khía cạnh này của thế giới nằm ngoài tầm với của khoa học vào thời điểm đó. Nhưng không phải là mãi mãi! Khi James Clerk Maxwell tìm ra phương trình đã được đặt theo tên ông vào năm 1861, mọi chuyện trở nên rõ ràng là ánh sáng và màu sắc cũng có thể hiểu được bằng toán học. Và giờ ta biết rằng phương trình Schrödinger nói đến ở trên, được phát hiện năm 1925, có thể dùng để dự đoán mọi đặc tính của vật chất, bao gồm cả độ cứng mềm. Trong khi tiến bộ về mặt lý thuyết đã đem lại ngày càng

nhiều dự đoán khoa học thì tiến bộ công nghệ đem lại ngày càng nhiều kiểm chứng thực nghiệm: gần như mọi thứ mà giờ ta nghiên cứu bằng kính viễn vọng, kính hiển vi hay máy gia tốc hạt đều nằm ngoài tầm với của khoa học trước đây. Nói cách khác, tầm hiểu biết của khoa học đã mở rộng đáng kể so với thời của Galileo, từ một tỷ lệ rất nhỏ đi đến được một phần lớn của tất cả các hiện tượng, bao gồm các hạt hạ nguyên tử, lỗ đen và nguồn gốc vũ trụ 13,8 tỷ năm trước. Điều này đặt ra câu hỏi: Vậy còn lại cái gì?

Đối với tôi, ý thức chính là rắc rối hiển nhiên mà không ai muốn nhắc đến. Bạn không chỉ biết rằng bạn có ý thức, mà đó còn là tất cả những gì bạn biết chắc chắn hoàn toàn – mọi thứ khác đều là do suy luận, như René Descartes đã chỉ ra từ thời Galileo. Liệu tiến bộ trong lý thuyết và công nghệ cuối cùng có đưa ý thức vào phạm trù khoa học vững chắc hay không? Ta không biết, cũng như Galileo không biết liệu có ngày con người hiểu được về ánh sáng và vật chất hay không.⁶³ Chỉ có một điều đảm bảo: ta sẽ không thành công nếu ta không thử! Đó là lý do tôi và nhiều nhà khoa học khác trên thế giới đang cố gắng hết mình để hình thành và kiểm chứng các lý thuyết về ý thức.

Các manh mối thực nghiệm về ý thức

Nhiều quá trình xử lý thông tin đang diễn ra trong đầu chúng ta ngay lúc này. Quá trình nào có ý thức và quá trình nào không? Trước khi tìm hiểu các lý thuyết ý thức và những gì chúng dự đoán, hãy xem các thí nghiệm từ trước đến nay đã cho ta biết điều gì, từ các quan sát truyền thống công nghệ thấp hay không cần công nghệ cho đến các đo đạc não tiên tiến nhất.

Hành vi nào có ý thức?

Nếu bạn nhân nhẩm 32 với 17 trong đầu, bạn ý thức được về nhiều phần của quá trình phép tính toán vận hành. Nhưng giả sử thay vào đó tôi cho bạn xem chân dung Albert Einstein và bảo bạn nói tên người đó. Như ta đã thấy trong chương 2, đây cũng là một nhiệm vụ tính toán: não bạn đang thực hiện một hàm số có đầu vào là thông tin về một số lượng lớn các màu sắc điểm ảnh gửi về từ mắt và đầu ra là thông tin gửi đến các cơ điều khiển miệng và dây thanh

quản. Các nhà khoa học máy tính gọi nhiệm vụ này là “phân loại hình ảnh” theo sau là “tổng hợp tiếng nói”. Mặc dù tính toán này phức tạp hơn nhiều so với nhiệm vụ nhân nhẩm, nhưng bạn lại có thể thực hiện nó nhanh hơn nhiều, gần như không cần cố gắng gì và không ý thức gì về chi tiết bạn làm điều đó

như thế nào.

Trải nghiệm chủ quan của bạn chỉ bao gồm nhìn vào bức hình, trải qua cảm giác nhận diện và rồi nghe mình nói ra “Einstein”.

Các nhà tâm lý học từ lâu đã biết ta có thể vô thức thực hiện một loạt các nhiệm vụ và hành vi khác, từ phản xạ chớp mắt đến hít thở, với tay, cầm nắm và giữ thăng bằng. Bạn thường nhận thức có ý thức về việc mình làm, nhưng không phải về cách mình làm. Mặt khác, các hành vi liên quan tới các tình huống không quen thuộc, tự chủ, các quy tắc logic phức tạp, lý luận trừu tượng hay xử lý ngôn ngữ thường là có ý thức. Chúng được gọi là tương quan hành vi của ý thức và có liên kết chặt chẽ với cách tư duy có cố gắng, chậm chạp và được kiểm soát mà các nhà tâm lý gọi là “Hệ thống 2”.⁵

Ta cũng biết rằng có thể chuyển đổi nhiều hành vi đều đặn từ trạng thái có ý thức sang trạng thái vô thức bằng cách tích cực thực hành, ví dụ như đi bộ, bơi, đạp xe, lái xe, đánh máy, cạo râu, buộc dây giày, chơi điện tử và chơi đàn piano.⁶ Quả thực, ai cũng biết rằng các chuyên gia thực hiện chuyên môn của mình giỏi nhất khi họ ở trong trạng thái “trôi chảy”, chỉ nhận thức được mọi thứ đang diễn ra ở một tầm cao hơn và không ý thức được những chi tiết cấp thấp như họ làm điều đó bằng cách nào. Ví dụ, hãy thử đọc câu tiếp theo trong tâm thái để ý đến từng chữ cái, giống như khi bạn mới tập đọc mà xem. Bạn có thấy làm như thế chậm hơn rất nhiều không, so với khi bạn chỉ ý thức về câu này ở cấp độ từ ngữ hay ý tưởng?

Trên thực tế, xử lý thông tin vô thức có vẻ không những khả thi mà còn giống quy tắc chung hơn là ngoại lệ. Ta đã có các chứng cứ cho thấy trong khoảng 107 bit thông tin đi vào não bộ mỗi giây từ các giác quan, ta chỉ nhận thức được một tỷ lệ rất nhỏ, ước tính từ 10

đến 50 bit.⁷ Điều này cho thấy quá trình xử lý thông tin mà ta nhận thức được rõ ràng chỉ là phần nổi của tảng băng trôi.

Kết hợp lại, những gợi ý này đã đưa một vài nhà nghiên cứu đến chỗ nhận định rằng quá trình xử lý thông tin có ý thức nên được coi như vị CEO của trí óc, chỉ xử lý những quyết định quan trọng nhất cần đến phân tích dữ liệu phức tạp từ mọi bộ phận của não bộ.⁸ Điều này giải thích vì sao, cũng như CEO trong một công ty, nó thường không muốn bị phân tâm vì biết mọi việc các cấp dưới đang làm – nhưng nó có thể tìm ra nếu muốn. Để trực tiếp trải nghiệm sự chú ý có chọn lọc này, hãy nhìn lại từ “muốn” một lần nữa: dán mắt vào dấu sắc, rồi không cử động đôi mắt, chuyển hướng chú ý từ dấu sắc ra các chữ cái và ra toàn bộ từ đó. Mặc dù thông tin từ võng mạc vẫn y nguyên, nhưng trải nghiệm có ý thức của bạn đã thay đổi. Cách ví von với CEO cũng giải thích tại sao công việc chuyên môn lại trở thành vô thức: sau khi đã vất vả để tìm ra được cách đọc và cách đánh máy, CEO giao lại những nhiệm vụ đều đặn này cho các cấp dưới vô thức để có thể tập trung vào những thách thức mới ở cấp độ cao hơn.

Ý thức nằm ở đâu?

Các thí nghiệm và phân tích khéo léo đã cho thấy ý thức bị giới hạn không chỉ trong một số hành vi nhất định, mà còn trong một số bộ phận nhất định của não bộ. Vậy đâu là nghi can chính? Nhiều gợi ý ban đầu đến từ các bệnh nhân bị chấn thương sọ não: những người có tổn thương cục bộ trong não do tai nạn, đột quỵ, khối u hay nhiễm trùng. Nhưng những quan sát này thường không mang tính kết luận. Ví dụ, liệu sự thật rằng chấn thương vùng phía sau não có thể gây mù có nghĩa là đây là khu vực của ý thức hình ảnh, hay chỉ có ý nghĩa đơn thuần là thông tin hình ảnh đi qua đó trên đường tới nơi mà nó sẽ trở thành ý thức, cũng giống như việc đầu tiên nó đi qua mắt?

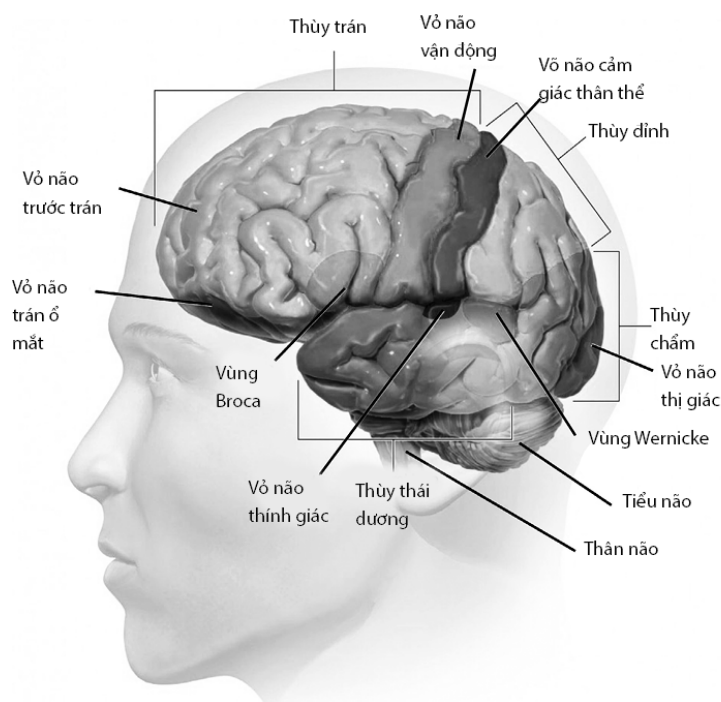
Mặc dù chấn thương và can thiệp y học không chỉ ra chính xác vị trí của các trải nghiệm ý thức, nhưng chúng cũng giúp ta thu hẹp lại các phương án. Ví dụ, tôi biết mặc dù tôi có cảm giác đau ở bàn tay như thể nó thật sự xảy ra ở đó, nhưng trải nghiệm cảm giác đau chắc chắn phải xảy ra ở một nơi khác, vì một nhà phẫu thuật từng giúp tôi

ngắt cảm giác đau bàn tay mà không hề đụng tới bàn tay tôi: ông chỉ cần gây tê các dây thần kinh ở vai tôi. Hơn nữa, một số người bị cắt bỏ tay trải qua ảo giác đau như thể nỗi đau đến từ bàn tay không còn tồn tại của họ. Một ví dụ khác, có đợt tôi để ý thấy mình chỉ nhìn bằng mắt phải, một phần trường thị lực của tôi bị thiếu mất – vậy là bác sĩ xác định võng mạc của tôi bị lỏng và đã gắn lại nó. Ngược lại, một số bệnh nhân chấn thương sọ não gặp phải tình trạng bán vô thị: họ cũng thiếu mất thông tin từ một nửa trường thị lực nhưng thậm chí không nhận thức được là nó bị thiếu – chẳng hạn như không để ý thấy mà ăn phần thức ăn ở nửa bên trái đĩa cơm. Khi ấy, như thể ý thức về một nửa thế giới của họ đã biến mất. Nhưng các vùng não bị tổn thương đó là nơi sinh ra trải nghiệm về không gian, hay chúng chỉ đơn thuần đưa thông tin về không gian đến nơi ý thức diễn ra, giống như võng mạc của tôi vậy?

Ở thập niên 1930, bác sĩ người Canada gốc Mỹ tiên phong trong ngành phẫu thuật thần kinh Wilder Penfield đã nhận thấy các bệnh nhân phẫu thuật thần kinh của ông báo cáo lại là các bộ phận khác nhau trên cơ thể họ được chạm vào khi ông kích điện những khu vực não nhất định ở chỗ ngày nay gọi là vùng cảm giác thân thể (hình 8.3).⁹ Ông cũng nhận thấy họ cử động các bộ phận cơ thể khác nhau một cách không tự nguyện khi ông kích thích các khu vực não ở chỗ ngày nay gọi là vùng vận động. Nhưng liệu điều đó có nghĩa là quá trình xử lý thông tin ở các khu vực não này tương ứng với ý thức về cảm giác chạm và vận động hay không?

May thay, công nghệ hiện đại giờ đây đem lại cho chúng ta những gợi ý chi tiết hơn nhiều. Dù ta vẫn còn rất xa mới đo được mọi xung điện phát đi từ khoảng hàng trăm tỷ nơ-ron, nhưng công nghệ đọc não đang tiến bộ rất nhanh, bao gồm các kỹ thuật với những cái tên đáng sợ như fMRI, EEG, MEG, ECoG, ePhys và cảm biến điện huỳnh quang. fMRI, viết tắt của functional magnetic resonance imaging (chụp cộng hưởng từ chức năng), đo đặc từ tính của các hạt nhân nguyên tử hydro để vẽ nên bản đồ 3-D của não bạn gần như mỗi giây, với độ phân giải từng milimet. EEG (điện não đồ) và MEG (từ não đồ) đo đặc điện trường và từ trường nằm bên ngoài đầu bạn để vẽ nên sơ đồ bộ não hàng ngàn lần mỗi giây, nhưng với độ phân

giải thấp, không có khả năng phân biệt các đặc điểm nhỏ hơn vài centimet. Nếu yếu bóng vía, bạn sẽ thích các kỹ thuật này vì chúng không cần đến phẫu thuật. Còn nếu không ngại bị mở hộp sọ, bạn có thêm các lựa chọn khác. Với ECoG (hay điện não đồ trong sọ), người ta đặt cả trăm dây điện lên bề mặt não, còn ePhys (điện sinh lý) lại bao gồm việc đưa các dây điện siêu vi, đôi khi mảnh hơn cả sợi tóc người, vào sâu trong não để ghi lại điện áp từ cả nghìn vị trí cùng một lúc. Nhiều bệnh nhân mắc chứng động kinh dành nhiều ngày trong bệnh viện trong khi người ta dùng ECoG để tìm ra bộ phận nào trong não đang gây ra các cơn co giật và nên được cắt bỏ. Họ cũng đồng ý để các nhà khoa học thần kinh thực hiện các thí nghiệm ý thức trên não họ. Cuối cùng, cảm biến điện huỳnh quang là phương pháp tác động di truyền lên nơ-ron khiến chúng phát ra tia sáng khi bắn xung điện, giúp cho hoạt động của nơ-ron có thể đo được bằng kính hiển vi. Trong số tất cả các kỹ thuật kể trên, kỹ thuật này có tiềm năng nhanh chóng theo dõi được số lượng nơ-ron lớn nhất, ít nhất là ở các loài động vật có não trong suốt – như loài giun *C. elegans* với 302 nơ-ron và ấu trùng cá ngựa vằn với khoảng 100.000 nơ-ron.



Hình 8.3: Các vùng thị giác, thính giác, cảm giác thân thể và vận động trên vỏ não lần lượt liên quan tới thị giác, thính giác, cảm giác chạm và kích hoạt cử động – nhưng điều đó không chứng tỏ chúng là nơi ý thức về việc nhìn, nghe, chạm và cử động diễn ra. Trên thực tế, nghiên cứu gần đây cho thấy vỏ não thị giác là hoàn toàn vô thức, cùng với tiểu não và thân não. Hình ảnh do Lachina (www.lachina.com) cung cấp.

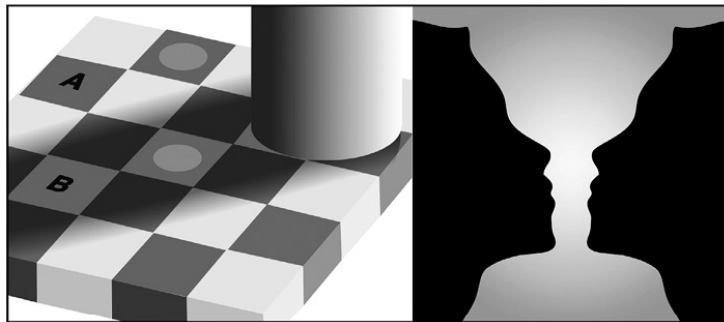
Mặc dù Francis Crick đã cảnh báo Christof Koch về chuyện nghiên cứu ý thức, nhưng Christof không chịu bỏ cuộc và cuối cùng còn thuyết phục được Francis. Năm 1990, họ viết bài nghiên cứu tiền đề về cái mà họ gọi là “tương quan thần kinh của ý thức” (NCC), đặt vấn đề quá trình não bộ cụ thể nào tương ứng với trải nghiệm ý thức nào. Trong hàng ngàn năm, các nhà tư duy đã tiếp cận với việc xử lý thông tin trong não bộ chỉ qua hành vi và trải nghiệm chủ quan của họ. Crick và Koch chỉ ra rằng công nghệ đọc não khi ấy đột nhiên cung cấp cách tiếp cận độc lập với các thông tin này, cho phép người ta nghiên cứu khoa học về vấn đề quá trình xử lý thông tin nào tương ứng với trải nghiệm ý thức nào. Hẳn nhiên, việc đo đạc nhờ công nghệ cho tới nay đã biến công cuộc tìm hiểu NCC thành một phần khá chính thống của khoa học thần kinh, với hàng nghìn bài viết xuất hiện trong những tạp chí khoa học danh tiếng nhất.¹⁰

Các kết luận đưa ra cho tới nay là gì? Để bước đầu cảm nhận công việc tìm hiểu NCC, hãy tự hỏi liệu võng mạc của bạn có ý thức hay chỉ là một hệ thống vô thức ghi lại thông tin hình ảnh, xử lý thông tin đó và gửi đến một hệ thống phía sau trong não bộ, nơi trải nghiệm chủ quan của bạn diễn ra. Trong khung hình trái của hình 8.4, hình vuông nào sẫm màu hơn: hình vuông mang chữ A hay chữ B? A phải không? Không phải, thật ra chúng có màu giống hệt nhau, bạn có thể kiểm chứng bằng cách nhìn từng hình vuông qua khe hở giữa các ngón tay. Điều đó chứng minh là trải nghiệm hình ảnh của bạn không thể nằm hoàn toàn trên võng mạc, vì nếu như vậy, hai hình vuông trông đã y hệt nhau.

Giờ nhìn vào khung hình bên phải của hình 8.4. Bạn nhìn thấy hai người phụ nữ hay một chiếc bình hoa? Nếu nhìn đủ lâu, bạn sẽ trải nghiệm chủ quan lần lượt cả hai hình ảnh đó, mặc dù thông tin đi tới

võng mạc của bạn vẫn y nguyên. Bằng cách đo đặc điều diễn ra trong não bộ trong hai tình huống đó, người ta có thể vạch rõ điều gì gây ra sự khác biệt – và đó không phải võng mạc, vì võng mạc phản ứng như nhau ở cả hai trường hợp.

Cú đấm khai tử giả thuyết võng mạc có ý thức đến từ một kỹ thuật gọi là “triệt tiêu chớp liên tục” do Christof Koch, Stanislas Dehaene và cộng sự đề xướng: họ phát hiện ra là nếu bạn để một trong hai mắt xem một chuỗi họa tiết phức tạp thay đổi liên tục, hệ thống thị giác của bạn sẽ bị phân tán tập trung tới mức bạn hoàn toàn không nhận thức được về hình ảnh tĩnh được đưa ra trước con mắt còn lại.¹¹ Tóm lại, bạn có thể có hình ảnh thị giác trên võng mạc mà không trải nghiệm được nó, và cũng có thể (trong khi mơ) trải nghiệm một hình ảnh không ở trên võng mạc của bạn. Điều này chứng tỏ hai bên võng mạc không phải là nơi ý thức thị giác diễn ra mà chỉ như chiếc máy quay, mặc dù chúng thực hiện những tính toán phức tạp liên quan đến hàng trăm triệu nơ-ron.



Hình 8.4: Hình vuông nào sẫm màu hơn – A hay B? Bạn nhìn thấy gì ở hình bên phải, một lọ hoa, hai người phụ nữ, hay lần lượt cả hai hình? Các ảo ảnh thị giác như thế này chứng tỏ ý thức thị giác không thể nằm ở mắt hay các giai đoạn ban đầu trong hệ thống thị giác, vì nó không chỉ phụ thuộc vào cái gì có trong bức tranh.

Các nhà nghiên cứu NCC cũng sử dụng phương pháp triệt tiêu chớp liên tục, ảo ảnh thị giác/thính giác không ổn định và các thủ thuật khác để chỉ ra các vùng nào trong não bộ có trách nhiệm với mỗi trải nghiệm ý thức của bạn. Chiến lược cơ bản là so sánh hoạt động của nơ-ron trong hai tình huống giống nhau về gần như mọi thứ (kể cả đầu vào giác quan) – ngoại trừ trải nghiệm ý thức của bạn. Các

bộ phận của não bộ hoạt động khác nhau (theo kết quả đo đạc) được xác định là các NCC.

Cách nghiên cứu NCC này đã chứng minh không có phần ý thức nào nằm trong bụng, cho dù đó là vị trí của hệ thống thần kinh ruột chứa tới 500 triệu nơ-ron tính toán cách để tối ưu hóa việc tiêu hóa thức ăn; còn các cảm giác như đói hay buồn nôn đều được sinh ra trong não bộ. Tương tự, không có phần ý thức nào có vẻ là nằm ở thân não – phần dưới cùng của não bộ kết nối với tủy sống, điều khiển hô hấp, nhịp tim và huyết áp. Kinh ngạc hơn nữa, ý thức có vẻ cũng không mở rộng đến tiểu não (hình 8.3), nơi chứa khoảng hai phần ba toàn bộ số lượng nơ-ron của bạn: các bệnh nhân bị tổn thương tiểu não thường gặp khó khăn khi nói và cử động vụng về gần như một kẻ say xỉn, nhưng vẫn hoàn toàn có ý thức.

Câu hỏi bộ phận nào của não bộ có trách nhiệm với ý thức vẫn còn để ngỏ và gây tranh cãi. Một số nghiên cứu NCC gần đây cho thấy ý thức chủ yếu nằm ở “vùng nóng” bao gồm đồi não (gần trung tâm não bộ) và phần phía sau của vỏ não (lớp ngoài cùng của não bộ, một tấm gồm sáu lớp có nhiều nếp nhăn, mà nếu trải phẳng ra sẽ có diện tích bằng một chiếc khăn ăn lớn).¹² Nghiên cứu này cũng đưa ra nhận xét gây tranh cãi rằng vỏ não thị giác nằm phía sau cùng của não bộ là một ngoại lệ, vì nó cũng vô thức như đôi mắt và võng mạc vậy.

Ý thức xảy ra khi nào?

Cho tới giờ, ta đã xem xét các gợi ý thực nghiệm về việc những loại xử lý thông tin nào có ý thức và ý thức diễn ra ở đâu. Nhưng còn câu hỏi ý thức diễn ra khi nào? Hồi nhỏ, tôi từng nghĩ là ta có ý thức về các sự kiện khi chúng xảy ra, mà tuyệt đối không có sự chậm trễ hay quãng thời gian nào ở giữa. Mặc dù đó vẫn là cảm giác chủ quan của tôi cho đến nay, nhưng rõ ràng điều đó không thể đúng được, vì chắc chắn não của tôi cần thời gian để xử lý thông tin đi vào qua các giác quan. Các nhà nghiên cứu NCC đã cẩn thận đo đạc thời gian này là bao lâu và kết luận của Christof Koch là cần khoảng một phần tư giây từ khi ánh sáng từ một vật thể phức tạp đi vào mắt bạn cho tới khi bạn nhìn thấy và nhận ra nó một cách có ý thức.¹³ Điều này nghĩa là nếu bạn đang đi trên đường cao tốc với tốc độ hơn 80km/h

và đột nhiên nhìn thấy một con sóc trước mặt bạn vài mét, thì đã quá muộn để làm gì được vì bạn đã chột qua nó mất rồi!

Tóm lại, ý thức của bạn sống trong thời quá khứ mà Christof Koch ước tính là chậm hơn thế giới bên ngoài khoảng một phần tư giây. Điều thú vị là bạn thường có thể phản ứng trước sự việc nhanh hơn bạn ý thức được chúng, chứng tỏ việc xử lý thông tin chịu trách nhiệm cho những phản ứng nhanh nhất của bạn phải là vô thức. Ví dụ, nếu một vật lạ tới gần mắt bạn, phản xạ chớp mắt có thể khép mắt bạn lại chỉ trong 1/10 giây. Điều này như thể một trong các hệ thống não bộ nhận được thông tin cảnh báo từ hệ thống thị giác, tính toán là mắt bạn có nguy cơ bị vật lạ đâm phải, gửi tín hiệu cho cơ mắt chớp mắt và đồng thời gửi tin nhắn cho phần có ý thức của bộ não rằng “Này, chúng ta sắp chớp mắt”. Đến lúc tin nhắn được đọc và đưa vào trải nghiệm ý thức của bạn thì cái chớp mắt đã diễn ra rồi.

Trên thực tế, hệ thống đọc tin nhắn đó liên tục bị dội bom bởi những tin nhắn đến từ khắp cơ thể, trong đó một số đến chậm hơn số khác. Để tín hiệu thần kinh đi từ ngón tay tới não cần nhiều thời gian hơn là từ mắt vì khoảng cách xa hơn, và phân tích hình ảnh mất nhiều thời gian hơn âm thanh vì nó phức tạp hơn – đó là lý do các cuộc đua Olympic bắt đầu bằng tiếng súng nổ chứ không phải bằng tín hiệu hình ảnh. Nhưng nếu chạm vào mũi, bạn trải nghiệm một cách có ý thức cảm giác trên mũi và trên đầu ngón tay như thể chúng xảy ra cùng lúc. Còn nếu vỗ tay, bạn nhìn thấy, nghe thấy và cảm thấy cái vỗ tay đó ở cùng thời điểm như nhau.¹⁴ Điều này nghĩa là trải nghiệm ý thức đầy đủ về một sự việc không được tạo ra cho tới khi tin nhắn báo cáo chậm chạp nhất đã đến nơi và được phân tích.

Một nhóm thí nghiệm NCC nổi tiếng do nhà sinh lý học Benjamin Libet tiên phong đã cho thấy loại hành động mà bạn có thể thực hiện vô thức không giới hạn ở các phản ứng nhanh như chớp mắt và đánh bóng bàn, mà còn bao gồm một số quyết định nhất định mà bạn có thể nghĩ là do ý chí tự do đưa ra – các đo đạc não bộ đôi khi có thể đoán trước quyết định của bạn trước cả khi bạn ý thức được là mình đã đưa ra nó.¹⁵

Các lý thuyết về ý thức

Như vừa thấy, mặc dù vẫn chưa hiểu rõ về ý thức nhưng ta có một lượng dữ liệu thực nghiệm đáng kể về nhiều khía cạnh của ý thức. Nhưng tất cả những dữ liệu này đến từ não bộ, vậy làm thế nào chúng có thể cho ta biết về ý thức của

máy móc?

Điều này đòi hỏi một phép ngoại suy vượt ra khỏi phạm vi thực nghiệm hiện tại. Nói cách khác, nó đòi hỏi một

lý thuyết.

Tại sao cần một lý thuyết?

Để hiểu được tại sao, hãy so sánh các lý thuyết về ý thức với các lý thuyết về trọng lực. Các nhà khoa học bắt đầu coi trọng lý thuyết trọng lực của Newton vì họ thu được từ đó nhiều hơn công sức mà họ đổ vào: những phương trình đơn giản chỉ vừa vặn một tờ giấy ăn có thể dự đoán chính xác kết quả của mọi thí nghiệm trọng lực từng được thực hiện. Do đó họ cũng coi trọng những dự đoán của lý thuyết này, vốn vượt ra khỏi phạm vi mà nó từng được kiểm chứng. Những phép ngoại suy táo bạo đó hóa ra lại áp dụng được cho cả chuyển động của các thiên hà trong những quần tụ trải ra hàng triệu năm ánh sáng. Tuy nhiên, các dự đoán này hơi sai lệch đối với chuyển động của Sao Thủy quanh Mặt Trời. Các nhà khoa học sau đó bắt đầu coi trọng thuyết tương đối đã được cải tiến của Einstein, hay thuyết tương đối rộng, vì có thể nói là nó thậm chí còn súc tích và ngắn gọn đủ ý hơn, lại dự đoán chính xác ngay cả điều mà lý thuyết của Newton đã sai. Kết quả là họ coi trọng tất cả các dự đoán vượt ra khỏi phạm vi đã được kiểm chứng của lý thuyết Einstein, về những hiện tượng xa xôi như lỗ đen, sóng hấp dẫn trong không thời gian và sự giãn nở của Vũ trụ từ một khởi đầu nóng bỏng tất cả những điều này sau đó đều được xác nhận bằng thực nghiệm.

Tương tự, nếu một lý thuyết toán học về ý thức với những phương trình vừa vặn trên một tờ giấy ăn có thể dự đoán thành công các kết quả của tất cả những thí nghiệm trên não bộ, thì ta sẽ coi

trọng không chỉ bản thân lý thuyết đó mà còn các dự đoán của nó về ý thức, vượt ra khỏi phạm vi của não bộ – trên máy móc chẳng hạn.

Ý thức nhìn từ quan điểm vật lý

Mặc dù một số lý thuyết về ý thức đã có từ thời cổ đại, hầu hết các lý thuyết hiện đại có nền tảng từ tâm lý học thần kinh và khoa học thần kinh, cố gắng giải thích và dự đoán về ý thức như những sự kiện thần kinh diễn ra trong não.¹⁶ Tuy các lý thuyết này đã đưa ra một số dự đoán thành công về tương quan thần kinh của ý thức, nhưng chúng không thể và cũng không có tham vọng đưa ra dự đoán về ý thức máy. Để thực hiện bước nhảy vọt từ não bộ sang máy móc, ta cần khái quát hóa từ NCC thành PCC:

tương quan vật lý của ý thức,

tức hình mẫu của các hạt chuyển động có ý thức. Bởi nếu một lý thuyết có thể dự đoán đúng cái gì có ý thức và cái gì không chỉ bằng những viên gạch vật lý cơ bản như các hạt cơ bản và các trường lực, thì nó có thể dự đoán không chỉ về não bộ mà còn về bất kỳ sự sắp xếp vật chất nào, kể cả các hệ thống AI tương lai. Vậy thì từ góc nhìn vật lý, cách sắp xếp hạt nào là có ý thức?

Nhưng câu hỏi này thật sự lại đặt ra một câu hỏi khác: Làm thế nào một thứ phức tạp như ý thức lại có thể tạo thành từ một thứ đơn giản như các hạt? Theo tôi thì vì đó là một hiện tượng có những đặc tính vượt lên trên và vượt ra khỏi đặc tính của các hạt. Trong vật lý học, ta gọi hiện tượng này là “đột sinh”.¹⁷ Hãy cùng tìm hiểu bằng cách xem xét một hiện tượng đột sinh đơn giản hơn ý thức: sự ứ đọng.

Một giọt nước ứ đọng, nhưng một tinh thể băng hay một đám mây hơi nước không ứ đọng, mặc dù chúng đều tạo thành từ các phân tử nước y hệt nhau. Tại sao lại như vậy? Vì đặc tính ứ đọng chỉ phụ thuộc vào cách sắp xếp phân tử. Không có nghĩa lý gì cả khi nói rằng một phân tử nước duy nhất ứ đọng, vì hiện tượng ứ đọng chỉ sinh ra khi có nhiều phân tử sắp xếp theo hình mẫu mà ta gọi là chất lỏng. Vậy chất rắn, chất lỏng và chất khí đều là các hiện tượng đột sinh: chúng nhiều hơn tổng các thành phần cộng lại, vì chúng có những đặc tính vượt lên trên và ra bên ngoài đặc tính của các hạt. Chúng có những đặc tính mà các hạt không có.

Giờ cũng như chất rắn, chất lỏng và chất khí, tôi nghĩ ý thức là hiện tượng đột sinh, với những đặc tính vượt lên trên và ra bên ngoài đặc tính của các hạt. Ví dụ, đi vào giấc ngủ sâu làm mất đi ý thức, bằng cách đơn giản là sắp xếp lại các hạt. Tương tự, ý thức của tôi sẽ biến mất nếu tôi chết cồng, khi đó các hạt tạo nên tôi được sắp xếp lại theo một cách kém may mắn hơn.

Khi bạn xếp nhiều hạt cùng với nhau để tạo thành bất cứ thứ gì từ nước cho đến não, các hiện tượng mới với những đặc tính quan sát được sẽ sinh ra. Giới vật lý chúng tôi thích nghiên cứu các đặc tính đột sinh này, thường có thể xác định bằng một tập hợp nhỏ các số mà bạn có thể đo được như độ kết dính của chất đó, độ chịu nén của chất đó, v.v. Ví dụ, nếu một chất kết dính đến mức rắn chắc, ta gọi đó là chất rắn, nếu không thì gọi đó là chất lưu. Và nếu chất lưu không nén lại được, ta gọi đó là chất lỏng, ngược lại thì gọi là chất khí hoặc chất plasma, tùy xem nó dẫn điện tốt đến đâu.

Ý thức như là thông tin

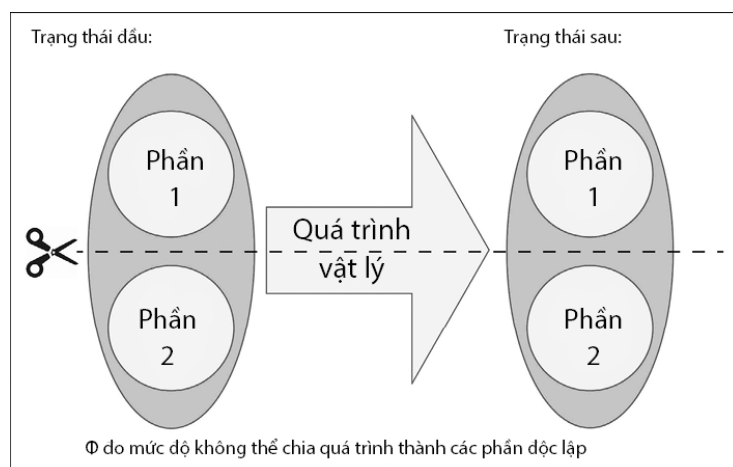
Vậy liệu có những số đo tương tự cho ý thức hay không? Nhà khoa học thần kinh người Ý Giulio Tononi đã đề xuất một số đo như thế, mà ông gọi là

“độ tích hợp thông tin”

, ký hiệu bằng chữ cái Hy Lạp Φ (

Phi

). Về cơ bản, nó đo mức độ mà các bộ phận khác nhau của một hệ thống biết về nhau (xem hình 8.5).



Hình 8.5: Với một quá trình vật lý biến đổi trạng thái ban đầu của hệ thống thành trạng thái mới theo thời gian, độ tích hợp thông tin Φ của nó đo mức độ không thể chia quá trình đó thành các phần độc lập. Nếu trạng thái tương lai của mỗi phần chỉ phụ thuộc vào quá khứ của nó mà không phụ thuộc vào những gì diễn ra ở phần kia thì $\Phi = 0$, nghĩa là cái mà ta gọi là hệ thống thật ra là hai hệ thống độc lập không hề có liên lạc với nhau.

Tôi gặp Giulio lần đầu tại một hội thảo vật lý năm 2014 ở Puerto Rico mà tôi đã mời ông cùng Christof Koch tham dự. Ấn tượng của tôi về ông là một nhà bác học đa tài chân chính tựa như Galileo và Leonardo da Vinci thời xưa. Đằng sau phong thái lặng lẽ là sự hiểu biết đến khó tin về nghệ thuật, văn học và triết học; ông còn có tiếng là đầu bếp giỏi: một phóng viên truyền hình từng kể với tôi rằng Giulio, chỉ trong vài phút, đã làm ra món salad ngon nhất anh từng được ăn. Tôi mau chóng nhận ra đằng sau cách ăn nói nhẹ nhàng là một trí thức không biết sợ hãi, quyết đi theo chứng cứ đến bất kỳ đâu nó dẫn ông đi, bất chấp những định kiến và cấm cản của giới chức quyền. Cũng như Galileo đã theo đuổi lý thuyết toán học về động lực mặc cho áp lực từ chính quyền là không được thách thức thuyết địa cầu là trung tâm, Giulio đã phát triển lý thuyết ý thức chính xác nhất về mặt toán học cho tới nay, thuyết thông tin tích hợp (IIT).

Tôi đã tranh luận suốt vài chục năm nay rằng ý thức là cách thông tin nảy sinh ra cảm giác khi được xử lý theo những cách phức tạp nhất định.¹⁸ IIT đồng quan điểm với lập luận này và thay thế cụm từ

mơ hồ “những cách phức tạp nhất định” của tôi bằng một định nghĩa chính xác và cụ thể: quá trình xử lý thông tin cần phải được tích hợp, tức là, số đo Φ phải lớn. Lập luận của Giulio vừa mạnh mẽ vừa đơn giản: hệ thống có ý thức cần được tích hợp thành một toàn thể thống nhất, vì nếu bao gồm hai phần độc lập thì nó sẽ có cảm giác như hai thực thể có ý thức tách biệt hơn là một. Nói cách khác, nếu một phần có ý thức của não bộ hay máy tính không thể liên lạc với phần còn lại, thì phần còn lại đó không thể là một phần trong trải nghiệm chủ quan của nó.

Giulio và các cộng sự đã đo đạc một phiên bản đơn giản hóa của Φ bằng cách sử dụng EEG để đo phản ứng của não bộ trước kích thích từ tính. “Máy phát hiện ý thức” của họ hoạt động rất tốt: nó xác định được rằng bệnh nhân có ý thức khi họ tỉnh táo hoặc đang nằm mơ, nhưng không có ý thức khi họ được gây mê hay đang ngủ say. Nó còn phát hiện ra ý thức ở hai bệnh nhân mắc hội chứng “khóa trong”, không thể cử động hay trò chuyện bằng bất kỳ cách thức thông thường nào.¹⁹ Vì vậy đây nổi lên như một công nghệ hứa hẹn cho các bác sĩ trong tương lai để xác định bệnh nhân có ý thức hay không.

Gắn liền ý thức với vật lý

IIT chỉ áp dụng được cho các hệ thống riêng biệt có số lượng trạng thái giới hạn, ví dụ như các bit trong bộ nhớ máy tính hay các nơ-ron đơn giản hóa với hai trạng thái bật hoặc tắt. Không may, điều này nghĩa là IIT không áp dụng được cho hầu hết các hệ thống vật lý truyền thống vốn có thể thay đổi liên tục – ví dụ, vị trí của một hạt hay độ mạnh của từ trường có thể mang bất kỳ giá trị nào trong một số lượng vô hạn các giá trị.²⁰ Nếu cố áp dụng công thức IIT cho những hệ thống đó, bạn thường nhận lại kết quả không mấy hữu ích là Φ bằng vô cực. Các hệ thống cơ học lượng tử có thể là riêng biệt, nhưng IIT ban đầu không áp dụng được cho các hệ thống lượng tử. Vậy làm thế nào để ta gắn liền IIT và các lý thuyết ý thức dựa trên thông tin khác với một nền tảng vật lý vững chắc?

Ta có thể làm vậy bằng cách phát triển từ những khám phá trong chương 2 về cách các khối vật chất có thể mang các đặc tính đột sinh liên quan tới thông tin. Ta đã thấy để một thứ dùng được như

thiết bị ghi nhớ có thể lưu trữ thông tin, nó cần có nhiều trạng thái bền vững khác nhau. Ta cũng thấy để trở thành computronium, một chất có thể thực hiện các tính toán, cũng cần cả những động lực phức tạp: các định luật vật lý cần làm nó thay đổi theo những cách đủ phức tạp để có thể thực hiện quá trình xử lý thông tin ngẫu nhiên. Cuối cùng, ta đã thấy một mạng nơ-ron là nền tảng mạnh mẽ cho việc học tập như thế nào vì chỉ cần đi theo các định luật vật lý là nó có thể tự sắp xếp chính mình để thực hiện ngày càng giỏi các tính toán mong muốn. Câu hỏi đặt thêm bây giờ là: Điều gì khiến một khối vật chất có thể có trải nghiệm chủ quan? Nói cách khác, trong các điều kiện nào thì một khối vật chất có thể làm bốn việc sau?

Ghi nhớ
Tính toán
Học tập
Trải nghiệm

Ta đã tìm hiểu về ba việc đầu ở chương 2 và giờ đang giải quyết đến việc thứ tư. Cũng như Margolus và Toffoli đã sáng tạo ra thuật ngữ computronium cho một chất có thể thực hiện các tính toán ngẫu nhiên, tôi muốn dùng thuật ngữ sentronium cho một chất chung nhất có trải nghiệm chủ quan (có tri giác).⁶⁴

Nhưng làm thế nào mà ý thức lại có được cảm giác phi vật lý như vậy nếu nó thật ra là một hiện tượng vật lý? Sao nó có thể có cảm giác độc lập với phần vỏ vật lý như vậy? Tôi nghĩ là vì nó thật sự khá độc lập với vỏ vật lý, thứ mà ở trong đó nó là một khuôn mẫu! Ta đã gặp nhiều ví dụ hay về các khuôn mẫu độc lập với vỏ vật lý trong chương 2, bao gồm sóng, ký ức và tính toán. Ta đã thấy chúng không chỉ nhiều hơn các bộ phận của chúng cộng lại (tức là mang tính đột sinh), mà còn độc lập với các bộ phận đó và có đời sống riêng. Ví dụ, ta đã thấy một trí óc mô phỏng hay nhân vật trò chơi điện tử trong tương lai sẽ không có cách nào biết được nó chạy trên hệ điều hành Windows, Mac OS, một chiếc điện thoại Android hay một hệ điều hành nào khác, vì nó độc lập với vỏ vật lý. Nó cũng không biết được các cổng logic của máy tính được làm từ transistor, vi mạch quang hay loại phần cứng nào khác. Chúng ta cũng đã thấy các định luật

vật lý cơ bản là gì – chúng có thể là bất kỳ thứ gì miễn sao chúng cho phép việc xây dựng máy tính phổ quát.

Tóm lại, tôi nghĩ ý thức là một hiện tượng vật lý có cảm giác phi vật lý vì giống như sóng và các tính toán, nó cũng có các đặc tính độc lập với vỏ vật lý cụ thể của nó. Điều này là sự tiếp nối hợp lý của ý tưởng cho rằng ý thức chính là thông tin. Nó còn dẫn đến một ý tưởng táo bạo mà tôi rất thích: Nếu ý thức là cách thông tin sinh ra cảm giác khi được xử lý theo những cách nhất định, thì nó phải độc lập với vỏ vật lý; thứ quan trọng là cấu trúc của việc xử lý thông tin chứ không phải là cấu trúc của vật chất đang thực hiện xử lý thông tin. Nói cách khác, ý thức độc lập với vỏ vật lý đến hai lần!

Như ta đã thấy, vật lý học mô tả các khuôn mẫu trong không thời gian tương ứng với các hạt chuyển động. Nếu các cách sắp xếp hạt tuân theo một số nguyên lý nhất định thì sẽ sinh ra các hiện tượng đột sinh khá độc lập với phần vỏ vật lý là các hạt và có cảm giác hoàn toàn khác. Một ví dụ hay chính là sự xử lý thông tin diễn ra trong các chất computronium. Nhưng giờ ta đã đưa ý tưởng này lên một tầm cao mới: Nếu bản thân sự xử lý thông tin tuân theo một số nguyên lý nhất định, thì nó có thể sinh ra một hiện tượng đột sinh ở cấp độ cao hơn mà ta gọi là ý thức. Như vậy trải nghiệm có ý thức nằm ở tầm cao hơn không phải một cấp mà là hai cấp so với vật chất. Thảo nào mà ý thức của bạn có cảm giác phi vật lý!

Điều này đặt ra câu hỏi: Sự xử lý thông tin cần tuân theo những nguyên lý nào để có ý thức? Tôi không giả bộ mình biết được những điều kiện đủ để đảm bảo cho ý thức là gì, nhưng sau đây là bốn điều kiện cần mà tôi dám cam đoan và quả thực đã khám phá được trong nghiên cứu của mình:

<i>Nguyên lý</i>	<i>Định nghĩa</i>
<i>Nguyên lý thông tin</i>	<i>Một hệ thống có ý thức có sức chứa thông tin lớn.</i>
<i>Nguyên lý động lực</i>	<i>Một hệ thống có ý thức có sức xử lý thông tin lớn.</i>
<i>Nguyên lý độc lập</i>	<i>Một hệ thống có ý thức có sự độc lập đáng kể với phần còn lại của thế giới.</i>
<i>Nguyên lý</i>	<i>Một hệ thống có ý thức không thể bao gồm các bộ</i>

tích hợp phận gần như độc lập với nhau.

Như đã nói, tôi nghĩ ý thức là cách thông tin nảy sinh ra cảm giác khi được xử lý theo những cách nhất định. Nghĩa là để có ý thức, một hệ thống cần có khả năng lưu trữ và xử lý thông tin, tức là cần hai nguyên lý đầu tiên. Lưu ý là bộ nhớ không cần kéo dài quá lâu: tôi khuyên bạn nên xem video cảm động về Clive Wearing, một người có vẻ hoàn toàn có ý thức dù ký ức của ông chỉ kéo dài chưa đầy một phút.²¹ Thêm nữa, tôi nghĩ là một hệ thống có ý thức cũng cần tương đối độc lập với phần còn lại của thế giới, vì nếu không nó sẽ không cảm thấy một cách chủ quan là nó có sự tồn tại độc lập nào. Cuối cùng, tôi nghĩ là hệ thống có ý thức cần tích hợp thành một toàn thể thống nhất, như Giulio Tononi lập luận, vì nếu nó bao gồm hai phần độc lập với nhau, thì chúng sẽ cảm thấy như hai thực thể có ý thức riêng biệt chứ không phải là một. Ba nguyên lý đầu tiên hàm ý về sự tự vận hành: hệ thống có khả năng lưu lại và xử lý thông tin mà không cần can thiệp nhiều từ bên ngoài, do đó tự quyết định tương lai của mình. Cả bốn nguyên lý hợp lại có ý nghĩa là hệ thống tự vận hành nhưng các bộ phận của nó thì không.

Nếu bốn nguyên lý này là đúng thì công việc của ta đã rõ ràng: ta cần tìm kiếm những lý thuyết vững chắc về mặt toán học có thể đại diện cho các nguyên lý đó và kiểm chứng chúng bằng thực nghiệm. Ta cũng cần xác định liệu có cần thêm các nguyên lý khác. Bất kể IIT có đúng hay không, các nhà nghiên cứu cũng nên cố gắng phát triển các lý thuyết cạnh tranh và kiểm chứng hết các lý thuyết hiện có bằng những thí nghiệm ngày càng tốt hơn.

Những tranh cãi về ý thức

Chúng ta đã bàn đến cuộc tranh luận dai dẳng về việc liệu nghiên cứu ý thức có phải là trò vớ vẩn không có tính khoa học và phí thời giờ vô ích hay không. Ngoài ra, gần đây còn có tranh cãi về những nghiên cứu ý thức tiên tiến nhất – hãy cùng tìm hiểu những tranh luận mà tôi thấy là giúp mở mang đầu óc nhất.

Lý thuyết IIT của Giulio Tononi gần đây đã thu hút không chỉ những lời khen ngợi mà cả phê phán, một số thậm chí còn khá gay gắt. Scott Aaronson mới đây viết trên blog cá nhân: “Theo ý tôi,

Thuyết Thông tin Tích hợp sai lầm thấy rõ vì những lý do từ cốt lõi, điều này đặt nó vào nhóm cỡ 2% những lý thuyết toán học hàng đầu về ý thức từng được đề xuất. Gần như tất cả các lý thuyết ý thức cạnh tranh nhau, như tôi thấy, đều thật mơ hồ, thiếu chắc chắn và dễ lung lay tới nỗi chỉ có thể dẫn đến sai lầm.”²² Trái chiều nhau là vậy, nhưng cả Scott và Giulio đều lịch sự lắng nghe lập luận của nhau khi tranh luận hùng hồn về IIT tại một lớp tập huấn ở Đại học New York gần đây mà tôi cũng tham dự. Aaronson chỉ ra rằng các mạng lưới cổng logic đơn giản nhất định có độ tích hợp thông tin (Φ) cực kỳ cao và lập luận là chúng rõ ràng không có ý thức, nên IIT là sai lầm. Giulio phản bác rằng nếu được xây dựng thì chúng sẽ có ý thức, và giả định của Scott rằng chúng không có ý thức là phiến diện, quá đề cao con người, chẳng khác gì một chủ lò mổ khẳng định động vật không thể có ý thức chỉ vì chúng không biết nói và rất khác con người. Và cả hai đều nhất trí với phân tích của tôi rằng họ bất đồng quan điểm về việc liệu tích hợp chỉ là điều kiện cần cho ý thức (mà Scott cho là đúng) hay còn là điều kiện đủ (như Giulio khẳng định). Khẳng định thứ hai rõ ràng là mạnh hơn và dễ gây bất đồng hơn, mà tôi hy vọng ta có thể sớm kiểm chứng bằng thực nghiệm.²³

Một khẳng định gây tranh cãi khác của IIT là thiết kế máy tính ngày nay không thể có ý thức, vì cách các cổng logic của chúng kết nối với nhau đem lại độ tích hợp rất thấp.²⁴ Nói cách khác, nếu bạn tải chính mình lên một robot trong tương lai có công nghệ cao và có thể mô phỏng chính xác từng nơ-ron và xi-nap của bạn, thì dù cho nhân bản số này có mang ngoại hình, cách nói năng và cử chỉ hành động giống hệt bạn đi nữa, Giulio khẳng định nó vẫn sẽ là một xác sống vô tri không có trải nghiệm chủ quan – khi ấy bạn sẽ hết sức thất vọng bởi bạn tải não lên vốn vì mong có được sự bất tử chủ quan.⁶⁵ Cả David Chalmers và giáo sư AI Murray Shanahan đều thách thức khẳng định này bằng cách hình dung điều gì sẽ xảy ra nếu bạn thay thế các mạch thần kinh trong não bộ bằng phần cứng kỹ thuật số giả định, có thể mô phỏng hoàn hảo các mạch này.²⁵ Mặc dù hành vi của bạn không bị ảnh hưởng bởi việc thay thế này vì sự mô phỏng đó được giả định là hoàn hảo, nhưng trải nghiệm của bạn sẽ thay đổi từ chỗ có ý thức lúc đầu sang không có ý thức khi quá

trình kết thúc, theo như lời Giulio nói. Nhưng cảm giác sẽ thế nào trong quá trình các mạch thần kinh được thay thế dần dần? Khi các bộ phận của não bộ chịu trách nhiệm cho trải nghiệm ý thức của nửa trường thị lực phía trên bị thay thế, bạn nhận ra một phần những gì mắt mình nhìn thấy đột nhiên bị thiếu đi, nhưng bằng cách bí ẩn nào đó bạn vẫn biết có gì ở đó, theo như ghi nhận của các bệnh nhân “thị lực mù” (những người bị khiếm thị do tổn thương não nhưng vẫn có khả năng phản ứng trước kích thích thị giác mà họ không nhìn thấy một cách có ý thức)?²⁶ Điều này sẽ gây hoang mang sâu sắc, vì nếu có thể trải nghiệm có ý thức bất kỳ sự khác biệt nào thì hẳn bạn cũng có thể kể cho bạn bè nghe điều đó khi được hỏi – nhưng theo giả định, hành vi của bạn lại không thể thay đổi. Khả năng có lý duy nhất phù hợp với các giả định này là chính tại thời điểm thứ gì đó biến mất khỏi ý thức của bạn, trí óc bạn được thay đổi một cách bí ẩn để hoặc là khiến bạn nói dối và phủ nhận rằng trải nghiệm của bạn thay đổi, hoặc quên mất là mọi thứ trước đây đã khác thế nào.

Mặt khác, Murray Shanahan thừa nhận cách phản biện tương tự dựa trên sự thay thế dần dần cũng có thể nhắm vào bất kỳ lý thuyết nào cho rằng ta có thể hành động như thể có ý thức mặc dù không có ý thức, nên bạn có thể đi đến kết luận là hành động và cảm giác có ý thức là một, và hành vi quan sát được từ bên ngoài do đó là tất cả những gì cần quan tâm. Nhưng khi ấy bạn lại rơi vào cái bẫy dự đoán rằng bạn không có ý thức khi đang nằm mơ, mặc dù bạn biết rõ là không phải vậy.

Tranh cãi thứ ba về IIT là liệu thực thể có ý thức có thể tạo thành từ các bộ phận có ý thức riêng biệt. Ví dụ, liệu toàn thể xã hội có thể có được ý thức mà không khiến từng người trong đó không mất đi ý thức hay không? Một bộ não có ý thức có thể có các bộ phận đồng thời tự ý thức về chính mình không? Dự đoán chắc chắn của IIT là “không”, nhưng không phải ai cũng thấy thuyết phục. Ví dụ, ở một số bệnh nhân có các chấn thương sọ não gây suy giảm nghiêm trọng liên lạc giữa hai bán cầu não xuất hiện “chứng tay ngoại lai”, nghĩa là bán cầu não phải khiến tay trái làm những việc mà bệnh nhân khẳng định là họ không cố tình làm và cũng không thể hiểu được – đôi khi đến mức họ phải dùng tay còn lại để ngăn chặn bàn tay “ngoại lai”.

Làm sao ta có thể chắc chắn rằng trong đầu họ không có hai ý thức tách biệt, một ở bán cầu não phải không có khả năng phát ngôn và một ở bán cầu não trái giành quyền đại diện phát ngôn cho cả hai? Hãy tưởng tượng ta sử dụng công nghệ tương lai để xây dựng một kết nối thông tin liên lạc trực tiếp giữa hai bộ não con người, rồi dần dần tăng cường kết nối này cho đến khi liên lạc giữa hai bộ não hiệu quả không khác gì liên lạc bên trong bộ não. Liệu có lúc nào hai ý thức cá nhân đó đột nhiên biến mất và bị thay thế bởi một ý thức thống nhất duy nhất như IIT dự đoán, hay sự chuyển đổi sẽ diễn ra từ từ để hai ý thức cá nhân cùng tồn tại ở dạng nào đó ngay cả khi một trải nghiệm chung bắt đầu nảy sinh?

Một tranh cãi thú vị khác là liệu các thí nghiệm có đánh giá thấp mức độ ý thức của chúng ta hay không. Ta đã thấy ở trên là mặc dù ta cảm thấy mình có ý thức thị giác về một lượng thông tin khổng lồ gồm màu sắc, hình khối, vật thể và gần như mọi thứ trước mắt ta, nhưng các thí nghiệm lại cho thấy ta chỉ nhớ được và ghi nhận lại một tỷ lệ nhỏ đến đáng buồn.²⁷ Một số nhà nghiên cứu đã cố gắng giải quyết sự chênh lệch này bằng cách đặt câu hỏi liệu đôi khi chúng ta có “ý thức không tiếp cận”, tức là có trải nghiệm chủ quan về những thứ quá phức tạp để nhét vừa vào trí nhớ ngắn hạn để dùng sau này.²⁸ Ví dụ, khi bạn trải qua trạng thái mù vô ý, tức là vì quá phân tâm nên không để ý tới một vật ngay trước mắt, điều này không có nghĩa là bạn không có trải nghiệm thị giác có ý thức về nó, mà chỉ là nó không được lưu trữ vào trí nhớ ngắn hạn của bạn.²⁹ Ta có nên coi đó là sự đánh tráo thay vì gọi là mù không? Các nhà nghiên cứu khác phản đối ý tưởng cho rằng không thể tin tưởng những gì người ta nói là người ta trải qua, đồng thời cảnh báo về các hàm ý của ý tưởng này. Murray Shanahan hình dung về một xét nghiệm lâm sàng trong đó bệnh nhân báo cáo là hoàn toàn không còn đau nhờ một loại thuốc kỳ diệu mới phát minh, nhưng sau đó bị một ban thẩm định của chính phủ bác bỏ: “Bệnh nhân chỉ tưởng là họ không đau thôi. Nhờ khoa học thần kinh, chúng ta biết rõ hơn họ.”³⁰ Mặt khác, đã có những trường hợp bệnh nhân vô tình tỉnh dậy giữa ca phẫu thuật được cho uống một loại thuốc để quên đi cảnh tượng khi đó. Ta có

nên tin khi sau đó họ báo cáo là mình không trải qua đau đớn gì không?³¹

Ý thức của AI cảm giác như thế nào?

Nếu một hệ thống AI tương lai có ý thức thì nó sẽ trải nghiệm chủ quan cái gì? Đây là cốt lõi của “vấn đề khó hơn” về ý thức, buộc ta đi lên tầng tiếp theo của độ khó như minh họa trong hình 8.1. Hiện tại ta không chỉ thiếu một lý thuyết trả lời được câu hỏi này, mà ta còn không chắc liệu có khả năng trả lời nó đầy đủ và hợp lý không. Suy cho cùng, một câu trả lời thỏa đáng sẽ như thế nào? Bạn sẽ giải thích về màu đỏ cho một người mù bẩm sinh như thế nào?

May thay, dù hiện tại chưa có câu trả lời đầy đủ nhưng điều đó cũng không ngăn được chúng ta đưa phần nào câu trả lời. Nếu có người ngoài hành tinh thông minh nghiên cứu hệ thống giác quan của con người, có lẽ họ sẽ suy luận màu sắc là cảm chất có cảm giác gắn liền với mỗi điểm trên một bề mặt hai chiều (trường thị lực), còn âm thanh không có cảm giác cục bộ về không gian như vậy, còn nỗi đau là cảm chất có cảm giác gắn liền với các bộ phận khác nhau trên cơ thể. Khi phát hiện võng mạc chúng ta có ba loại tế bào nón nhạy cảm với ánh sáng, họ có thể suy ra rằng chúng ta trải nghiệm được ba màu cơ bản và các cảm chất màu sắc khác đều là kết quả của việc kết hợp ba màu này. Bằng cách đo đặc thời gian cần để nơ-ron truyền thông tin khắp não bộ chúng ta, họ có thể kết luận ta trải nghiệm không quá 10 ý nghĩ hay cảm nhận có ý thức mỗi giây, và khi ta xem phim trên TV với tốc độ 24 hình trên giây, cái ta trải nghiệm không phải một chuỗi hình ảnh tĩnh mà là chuyển động liên tục. Khi đo đặc tốc độ adrenaline được giải phóng vào mạch máu chúng ta và thời gian nó giữ nguyên trước khi bị phá vỡ, họ có thể dự đoán ta cảm thấy cơn giận dâng trào bắt đầu chỉ trong vài giây và kéo dài nhiều phút.

Áp dụng những lập luận dựa trên vật lý học tương tự, ta có thể đưa ra dự đoán có cơ sở về một vài khía cạnh nhất định của cảm giác mà ý thức nhân tạo trải qua. Đầu tiên, phạm vi trải nghiệm có thể có của AI là vô cùng lớn so với những gì con người có thể trải nghiệm. Ta có một loại cảm chất cho mỗi loại giác quan, nhưng số dạng cảm biến và đại diện nội bộ cho thông tin của AI có thể lớn hơn

rất nhiều, nên ta phải tránh cái bẫy giả định rằng cảm giác khi là một AI thì nhất thiết phải tương tự như cảm giác khi là người.

Thứ hai, số trải nghiệm mỗi giây của một ý thức nhân tạo kích cỡ bộ não có thể lớn hơn ta hàng triệu lần, vì các tín hiệu điện từ truyền đi với tốc độ ánh sáng – nhanh hơn hàng triệu lần so với tín hiệu thần kinh. Tuy nhiên, AI càng lớn, các ý nghĩ toàn thể của nó càng phải chậm lại để có thời gian cho thông tin truyền đi khắp các bộ phận, như ta đã thấy trong chương 4. Do đó ta có thể ước tính một AI “Gaia” kích cỡ Trái Đất chỉ có khoảng 10 trải nghiệm ý thức mỗi giây, cũng như con người, và một AI kích cỡ thiên hà thì cứ mỗi 100.000 năm mới có được một ý nghĩ – vậy là không quá 100 trải nghiệm trong toàn bộ lịch sử Vũ trụ của ta cho tới nay! Đây là một động cơ dằn vặt như không thể cưỡng lại, khiến các AI lớn phải giao phó công việc tính toán cho những hệ thống con nhỏ nhất có thể xử lý được chúng, để tăng tốc, cũng như trí óc có ý thức của ta đã giao phó phản xạ chớp mắt cho một hệ thống con nhỏ bé, nhanh nhẹn và vô thức. Dù đã thấy ở trên là xử lý thông tin có ý thức trong não bộ có vẻ chỉ là phần nổi của tảng băng trôi và phần còn lại thì không có ý thức, nhưng ta nên kỳ vọng tình hình còn cực đoan hơn đối với các AI lớn trong tương lai: nếu chúng có một ý thức duy nhất thì khả năng chúng sẽ không nhận thức được hầu hết các quá trình xử lý thông tin diễn ra bên trong ý thức này. Hơn nữa, các trải nghiệm ý thức mà nó tận hưởng có thể cực kỳ phức tạp nhưng cũng chỉ mang tốc độ rùa bò so với hoạt động nhanh chóng của các bộ phận nhỏ hơn.

Điều này đưa đến đỉnh điểm tranh cãi đã nhắc đến ở trên về việc liệu các bộ phận của thực thể có ý thức có thể đồng thời có ý thức hay không. IIT dự đoán là không, nghĩa là nếu một AI to lớn trong vũ trụ tương lai có ý thức thì hầu hết các quá trình xử lý thông tin của nó đều vô thức. Điều này tức là nếu một nền văn minh gồm các AI nhỏ hơn cải thiện khả năng liên lạc đến mức sinh ra một trí óc tổ hợp có ý thức duy nhất, thì các ý thức cá nhân nhanh hơn nhiều bỗng nhiên bị loại bỏ. Tuy nhiên nếu dự đoán của IIT là sai, trí óc tổ hợp có thể cùng tồn tại với rất nhiều trí óc có ý thức nhỏ hơn. Người ta thậm chí có thể tưởng tượng ra một bậc thang ý thức lồng ghép vào nhau ở mọi cấp độ từ hiển vi tới vũ trụ.

Như ta đã thấy ở trên, việc xử lý thông tin vô thức trong não người có vẻ liên quan đến cách tư duy nhanh chóng, tự động và không cần cố gắng mà các nhà tâm lý học gọi là “Hệ thống 1”.³² Ví dụ, Hệ thống 1 của bạn có thể thông báo với ý thức là phân tích cực kỳ phức tạp của nó dựa trên dữ liệu đầu vào từ thị giác đã xác định là bạn thân của bạn đã tới, mà không cho bạn biết chút gì về cách tính toán đó diễn ra. Nếu mối liên kết này giữa hệ thống và ý thức được xác thực, thì “Hệ thống 1” trở thành cách khái quát hóa hấp dẫn cho các AI, đại diện cho tất cả các nhiệm vụ lặp lại đều đặn nhanh chóng được giao cho các đơn vị cấp dưới vô thức là Hệ thống 1 của AI. Còn tư duy toàn thể chậm chạp, có cố gắng và được kiểm soát của AI sẽ, nếu có ý thức, được gọi là Hệ thống 2 của AI. Con người chúng ta còn có trải nghiệm ý thức liên quan đến cái mà tôi sẽ gọi là “Hệ thống 0”: những nhận thức thô sơ diễn ra bị động ngay cả khi bạn đang ngồi yên không cử động hay suy nghĩ mà chỉ quan sát thế giới xung quanh. Hệ thống 0, 1 và 2 dường như có độ phức tạp tăng dần theo thứ tự, nên thật đáng kinh ngạc là chỉ hệ thống ở giữa là có vẻ vô thức. IIT cho rằng điều này là do thông tin giác quan thô sơ ở Hệ thống 0 được lưu trữ trong những cấu trúc não bộ như bàn cờ có độ tích hợp rất cao, còn Hệ thống 2 có độ tích hợp cao nhờ các vòng phản hồi, nơi tất cả thông tin bạn nhận thức được ngay lúc này có thể ảnh hưởng đến trạng thái não bạn trong tương lai. Mặt khác, chính dự đoán về “bàn cờ ý thức” đã khiến IIT nhận phải sự chỉ trích của Scott Aaronson như đã nói ở trên. Tóm lại, nếu ngày nào đó một lý thuyết giải quyết vấn đề khá khó của ý thức có thể vượt qua một chuỗi thí nghiệm kiểm chứng khắt khe để rồi chúng ta bắt đầu nhìn nhận nghiêm túc các dự đoán của nó, thì nó cũng sẽ thu hẹp lại đáng kể các phương án cho vấn đề còn khó hơn là AI có ý thức trong tương lai có thể trải nghiệm điều gì.

Một số khía cạnh trong trải nghiệm chủ quan của chúng ta rõ ràng có nguồn gốc từ quá trình tiến hóa, chẳng hạn như những ham muốn cảm xúc liên quan đến việc tự bảo toàn (ăn, uống, tránh bị giết) và sinh sản. Điều này có nghĩa là có thể tạo ra AI không bao giờ trải nghiệm các cảm chất như đói, khát, sợ hay ham muốn tình dục. Như ta đã thấy trong chương trước, nếu một AI cực kỳ thông minh được

lập trình để có gần như bất kỳ mục tiêu đủ tham vọng nào, khả năng cao là nó sẽ nỗ lực tự bảo toàn để đạt được mục tiêu đó. Tuy nhiên, nếu là một phần của một xã hội các AI thì những AI này có thể thiếu đi nỗi sợ chết mạnh mẽ của con người: một khi đã có một bản sao lưu dự phòng thì bất quá chúng chỉ mất đi các ký ức đã tích lũy từ lần sao lưu gần đây nhất, miễn là chúng tự tin phần mềm sao lưu của mình sẽ được sử dụng. Thêm vào đó, khả năng sẵn sàng sao chép thông tin và phần mềm giữa các AI có lẽ sẽ làm giảm ý thức cá nhân mạnh mẽ vốn là đặc điểm nổi trội của ý thức con người: giữa bạn và tôi sẽ có ít sự phân biệt nếu chúng ta có thể dễ dàng chia sẻ hay sao chép các ký ức và khả năng của chúng ta, nên một nhóm AI ở gần nhau sẽ có cảm giác như một sinh vật duy nhất mang một trí óc tổ hợp.

Liệu ý thức nhân tạo có cảm giác rằng nó có ý chí tự do hay không? Lưu ý rằng, mặc dù các nhà triết học đã mất hàng thiên niên kỷ cãi cọ về việc chúng ta có ý chí tự do hay không mà vẫn chưa thể thống nhất thậm chí từ cách định nghĩa câu hỏi này,³³ tôi đang đặt ra một câu hỏi khác, có thể nói là dễ giải quyết hơn. Để tôi thử thuyết phục bạn rằng câu trả lời đơn giản là “Có, bất kỳ thực thể nào đưa ra quyết định có ý thức đều sẽ chủ quan cảm thấy rằng mình có ý chí tự do, bất kể đó là thực thể sinh học hay nhân tạo.” Các quyết định nằm trên một phổ giữa hai thái cực:

Bạn biết chính xác tại sao bạn đưa ra lựa chọn cụ thể đó.

Bạn hoàn toàn không biết tại sao bạn đưa ra lựa chọn cụ thể đó
– như thể bạn chỉ lựa chọn ngẫu hứng.

Các cuộc thảo luận về ý chí tự do thường xoay quanh sự vật lộn để khiến hành vi đưa ra quyết định hướng tới mục tiêu hòa hợp với các định luật vật lý: nếu bạn đang lựa chọn giữa hai lời giải thích sau về việc mình làm thì cái nào đúng: “Tôi hẹn cô ấy đi chơi vì tôi thật sự thích cô ấy” hay “Các hạt của tôi khiến tôi làm thế bằng cách chuyển động theo các định luật vật lý”? Nhưng ta đã thấy trong chương trước là cả hai đều đúng: cái tạo ra cảm giác như hành vi hướng tới mục tiêu có thể nảy sinh từ các định luật vật lý tất định không có mục tiêu gì. Cụ thể hơn, khi một hệ thống (não bộ hay AI) đưa ra quyết

định loại 1, nó sử dụng một số thuật toán tắt định để tính toán cần quyết định điều gì, và cảm giác cho rằng tự nó quyết định là kết quả của thực tế là nó đã quyết định khi tính toán cần phải làm gì. Hơn nữa, như Seth Lloyd đã nhấn mạnh,³⁴ có một định lý khoa học máy tính nổi tiếng nói rằng đối với hầu hết các tính toán, không có cách nào xác định kết quả nhanh hơn là chạy các tính toán đó. Điều này nghĩa là thông thường bạn không thể nghĩ ra trong thời gian ít hơn một giây là mình sẽ quyết định làm gì một giây sau đó, điều này giúp củng cố trải nghiệm là bạn có ý chí tự do. Ngược lại, khi một hệ thống (não bộ hay AI) đưa ra một quyết định loại 2, nó chỉ đơn giản là lập trình trí óc mình để đưa ra quyết định dựa trên kết quả đầu ra của một hệ thống con có vai trò như máy tạo số ngẫu nhiên. Trong não bộ và máy tính, các con số ngẫu nhiên được sinh ra dễ dàng bằng cách tăng cường độ ồn. Do đó, bất kể quyết định rơi vào khoảng nào trên phổ từ 1 đến 2, cả ý thức sinh học lẫn nhân tạo đều cảm thấy nó có ý chí tự do: nó cảm giác thực sự chính nó là người quyết định và không thể đoán chắc quyết định sẽ là gì cho tới khi nó đã suy nghĩ xong.

Vài người nói với tôi rằng họ thấy quan hệ nhân quả hạ thấp giá trị của họ, khiến quá trình tư duy của họ trở nên vô nghĩa còn họ chỉ là những cỗ máy “đơn thuần”, không hơn. Tôi thấy thái độ tiêu cực như vậy thật kỳ cục và vô căn cứ. Đầu tiên, bộ não con người chẳng có gì là “đơn thuần” cả, thậm chí tôi còn thấy đó là vật thể vật lý phức tạp kỳ diệu nhất trong Vũ trụ đã biết của ta. Thứ hai, vậy thì họ muốn giải pháp thay thế nào? Chẳng phải họ muốn chính quá trình tư duy của họ (những tính toán do não họ thực hiện) đưa ra quyết định hay sao? Trải nghiệm chủ quan của họ về ý chí tự do chỉ là cách các tính toán của họ cảm giác từ bên trong: họ không biết kết quả của một tính toán cho tới khi họ tính toán xong. Đó là ý nghĩa của câu nói “tính toán chính là quyết định”.

Ý nghĩa

Hãy kết thúc bằng cách trở lại điểm bắt đầu của cuốn sách: Chúng ta muốn tương lai sự sống trở nên như thế nào? Trong chương trước, ta đã thấy các nền văn hóa đa dạng trên toàn cầu đều

tìm kiếm một tương lai hứa hẹn đầy trải nghiệm tích cực, nhưng nhiều tranh cãi gai góc hết sức thú vị lại nảy sinh khi người ta tìm cách nhất trí đâu là cái nên được coi là tích cực và làm thế nào để thỏa hiệp giữa những thứ tốt cho các dạng sống khác nhau. Nhưng đừng để những tranh cãi đó làm ta phân tán tập trung khỏi vấn đề rõ mồn một nhưng thường bị né tránh: không thể có trải nghiệm tích cực nếu như trải nghiệm không tồn tại, tức là nếu ý thức không tồn tại. Nói cách khác, nếu không có ý thức, sẽ không có niềm hạnh phúc, lòng tốt, cái đẹp, ý nghĩa hay mục đích mà chỉ có sự phí phạm không gian trên toàn vũ trụ. Điều này hàm ý là khi người ta hỏi về ý nghĩa cuộc sống như thể nhiệm vụ của vũ trụ là đem lại ý nghĩa cho sự tồn tại của chúng ta, họ đang hiểu theo chiều ngược lại:

Không phải Vũ trụ của ta đem lại ý nghĩa cho các sinh vật có ý thức, mà chính các sinh vật có ý thức đem lại ý nghĩa cho Vũ trụ của ta.

Vậy mục tiêu đầu tiên trong danh sách điều ước cho tương lai nên là duy trì (và hy vọng mở rộng) ý thức sinh học và/hoặc nhân tạo trong vũ trụ, thay vì để nó tuyệt chủng.

Nếu ta thành công trong công cuộc này thì con người chúng ta sẽ cảm thấy thế nào về việc cùng tồn tại bên những cỗ máy ngày càng thông minh? Sự trỗi dậy dường như không thể lay chuyển của trí tuệ nhân tạo có làm bạn thấy khó chịu không, và nếu có thì tại sao? Trong chương 3, ta đã thấy công nghệ chạy bằng AI có thể thỏa mãn các nhu cầu cơ bản của chúng ta tương đối dễ dàng, như an sinh và thu nhập, miễn là ý chí chính trị để làm điều đó có tồn tại. Tuy nhiên, có thể bạn lo ngại rằng được ăn no mặc ấm và an cư lạc nghiệp là chưa đủ. Nếu ta được đảm bảo là AI sẽ chăm sóc tất cả các nhu cầu và ham muốn thực tiễn của chúng ta, thì liệu cuối cùng ta có thấy cuộc đời mình thiếu đi ý nghĩa và mục đích, giống như thú vật được vỗ béo trong sở thú hay không?

Từ xưa đến nay, con người chúng ta vẫn thường xây dựng giá trị bản thân dựa trên ý tưởng con người là bá chủ: niềm tin rằng chúng ta là những sinh vật thông minh nhất hành tinh, độc đáo và tối thượng. Sự trỗi dậy của AI sẽ buộc chúng ta từ bỏ ý tưởng này và

trở nên khiêm tốn hơn. Nhưng dù sao có lẽ đó cũng là điều ta nên làm: suy cho cùng, bám víu lấy ý tưởng độc tôn rằng ta là bậc bề trên (so với cá nhân khác, nhóm sắc tộc khác, loài khác, v.v.) đã gây ra những vấn đề tồi tệ khủng khiếp trong quá khứ và là một ý tưởng có lẽ đã hết thời. Quả thực, coi con người là bá chủ không chỉ gây ra đau thương trong quá khứ mà còn có vẻ không cần thiết cho sự sinh sôi phát triển của loài người: nếu ta phát hiện thấy một nền văn minh hòa bình ngoài Trái Đất, tiên tiến hơn ta nhiều về mặt khoa học, nghệ thuật và mọi thứ khác mà ta quan tâm, hẳn điều đó cũng sẽ không ngăn cản người ta tiếp tục trải nghiệm ý nghĩa và mục đích trong cuộc sống. Ta có thể duy trì gia đình, bạn bè, các cộng đồng lớn hơn cùng tất cả các hoạt động đem lại cho ta ý nghĩa và mục đích, và hy vọng không mất đi thứ gì khác ngoài sự tự kiêu.

Khi ta lên kế hoạch cho tương lai, hãy cân nhắc không chỉ ý nghĩa của cuộc sống chúng ta mà cả ý nghĩa của bản thân Vũ trụ. Ở điểm này, hai nhà vật lý tôi yêu quý là Steven Weinberg và Freeman Dyson lại đại diện cho hai quan điểm hoàn toàn trái ngược. Weinberg, người giành giải Nobel cho công trình đặt nền móng về mô hình chuẩn của vật lý hạt, có câu nói nổi tiếng “Vũ trụ càng trở nên dễ hiểu với chúng ta thì dường như càng tỏ ra vô nghĩa.”³⁵ Trái lại, Dyson lạc quan hơn nhiều, như ta đã thấy trong chương 6: dù đồng ý là Vũ trụ của ta đã từng vô nghĩa, nhưng ông tin rằng sự sống giờ đây đang tự lấp đầy với ngày càng nhiều ý nghĩa hơn, trong đó ý nghĩa tuyệt vời nhất sẽ đến nếu sự sống lan tỏa thành công khắp vũ trụ. Ông kết thúc bài nghiên cứu tiên phong năm 1979 như sau: “Liệu vũ trụ của Weinberg hay của tôi gần với sự thật hơn? Một ngày nào đó, không lâu nữa, chúng ta sẽ biết.”³⁶ Nếu Vũ trụ của ta trở lại trạng thái vô thức vĩnh viễn vì ta để sự sống trên Trái Đất tuyệt chủng hay vì ta để AI xác sống vô thức chiếm trọn Vũ trụ, thì đó sẽ là bằng chứng hùng hồn chứng minh Weinberg đúng.

Từ quan điểm này, ta thấy rằng dù tương lai sự sống vẫn là tiêu điểm trong cuốn sách này, nhưng tương lai của ý thức còn quan trọng hơn nữa vì đó là điều đem lại ý nghĩa cho sự sống. Các triết gia thích dùng tiếng Latin để phân biệt sự đối lập giữa sapience (khả năng suy nghĩ thông minh) với sentience (khả năng trải nghiệm cảm

chất một cách chủ quan). Con người chúng ta đã xây dựng nhân dạng cho mình từ nền tảng Homo sapiens, những sinh vật thông minh nhất trên đời. Khi chúng ta chuẩn bị phải nhún nhường trước những máy móc ngày càng thông minh hơn, tôi gợi ý chúng ta nên xây dựng lại thương hiệu khác cho bản thân, là Homo sentiens!

TÓM TẮT:

Không có định nghĩa chuẩn chỉ nào về “ý thức”. Tôi dùng định nghĩa rộng và không đặt con người làm trung tâm là ý thức = trải nghiệm chủ quan.

Việc AI có ý thức theo nghĩa đó hay không là điều quan trọng đối với những vấn đề triết học và đạo đức gai góc nhất mà sự trỗi dậy của AI đặt ra: AI có chịu đau khổ không? AI có nên có quyền không? Tải nào lên có phải là tự sát trên phương diện chủ quan không? Vũ trụ tương lai đầy rẫy AI có thể nào là tận thế xác sống thành hiện thực không?

Vấn đề hiểu được trí tuệ không nên bị đánh đồng với ba vấn đề riêng biệt của ý thức: “vấn đề khó” là dự đoán hệ thống vật lý nào có ý thức, “vấn đề khó hơn” là dự đoán về cảm chất, và “vấn đề khó thật” là tại sao lại có ý thức.

“Vấn đề khó” của ý thức là vấn đề mang tính khoa học, vì ta có thể kiểm chứng và xác định bằng thực nghiệm tính đúng-sai của một lý thuyết dự đoán các quá trình nào trong não bộ có ý thức, trong khi vẫn chưa rõ làm thế nào mà khoa học có thể giải quyết đầy đủ hai vấn đề khó hơn.

Các thí nghiệm khoa học thần kinh cho thấy nhiều hành vi và khu vực não bộ không có ý thức, mà phần lớn trải nghiệm ý thức là đại diện cho một bản tóm tắt sau khi sự việc đã xảy ra của một lượng thông tin vô thức lớn hơn nhiều.

Khái quát hóa các dự đoán về ý thức từ não bộ sang máy móc cần đến một lý thuyết. Ý thức dường như không đòi hỏi một loại hạt hay trường đặc biệt nào, mà cần một loại xử lý thông tin đặc biệt mang tính tự vận hành và tích hợp tương đối, sao cho cả hệ thống là tự vận hành nhưng các bộ phận của hệ thống thì không.

Ý thức có thể có cảm giác phi vật lý như vậy vì nó hai lần độc lập với vỏ vật lý: nếu ý thức là cách thông tin sinh ra cảm giác khi được xử lý theo những cách phức tạp nhất định, thì điều quan trọng là cấu trúc của sự xử lý thông tin chứ không phải cấu trúc của vật chất thực hiện xử lý thông tin.

Nếu ý thức nhân tạo là khả thi, thì phạm vi của các trải nghiệm AI khả thi là rất lớn so với những gì con người có thể trải nghiệm, trải rộng trên một phổ cảm chất và quy mô thời gian lớn – tất cả cùng chia sẻ một cảm giác là bản thân có ý chí tự do.

Vì không thể có ý nghĩa nếu không có ý thức, nên không phải Vũ trụ của ta đem lại ý nghĩa cho các sinh vật có ý thức, mà chính sinh vật có ý thức đem lại ý nghĩa cho Vũ trụ của ta.

Điều này cho thấy khi con người chúng ta chuẩn bị để khiêm nhường trước những máy móc ngày càng thông minh, ta sẽ khuây khỏa khi nghĩ mình là Homo sentiens, không phải Homo sapiens.

Phần kết

Câu chuyện của đội FLI

Điều buồn nhất của cuộc sống lúc này là khoa học thu được hiểu biết nhanh hơn xã hội thu được sự thông tuệ.

Isaac Asimov

Bạn đọc thân mến, vậy là chúng ta đang ở đây, chặng cuối của cuốn sách, sau khi đã khám phá về nguồn gốc và số phận của trí tuệ, về mục tiêu và ý nghĩa. Vậy làm thế nào để ta có thể đưa những ý tưởng ấy thành hành động? Chúng ta nên làm điều cụ thể gì để xây dựng tương lai tốt đẹp nhất có thể? Đây chính là câu hỏi tôi đang tự hỏi mình lúc này, khi ngồi đây trên ghế sắt cửa sổ trong chuyến bay từ San Francisco về lại Boston ngày 9 tháng 11 năm 2017, từ hội nghị AI chúng tôi vừa tổ chức ở Asilomar. Bởi vậy, hãy để tôi kết thúc cuốn sách bằng việc chia sẻ những suy nghĩ của tôi với các bạn.

Cạnh tôi, Meia đang tranh thủ ngủ sau nhiều đêm chỉ ngủ được chút ít vì còn phải chuẩn bị tổ chức hội nghị. Chà – một tuần thật náo động! Chúng tôi đã mời được gần như tất cả những người tôi đã đề cập trong sách này tới gặp nhau trong vài ngày ở phần tiếp theo này của hội nghị Puerto Rico, trong đó có các doanh nhân như Elon Musk và Larry Page cùng các nhà nghiên cứu AI tiên phong từ các trường đại học và các công ty như DeepMind, Google, Facebook, Apple, IBM, Microsoft và Baidu, cũng như các nhà kinh tế, luật gia, triết gia và các nhà tư tưởng tuyệt vời khác (xem hình 9.1). Kết quả vượt cả những mong đợi cao nhất của tôi, và tôi thấy lạc quan hơn bao giờ hết về tương lai sự sống. Trong phần kết này, tôi sẽ nói cho bạn biết vì sao.

FLI ra đời

Kể từ khi biết về cuộc chạy đua vũ trang hạt nhân ở tuổi 14, tôi đã lo ngại rằng sức mạnh của công nghệ đang tăng lên nhanh hơn sự thông thái mà ta có để quản lý nó. Do đó tôi quyết định dành một chương trong cuốn sách đầu tay của tôi, Vũ trụ toán học của ta, để

nói về thách thức này, dù phần còn lại chủ yếu là về vật lý học. Tôi đặt ra quyết tâm cho năm 2014 là sẽ không được than thở về bất cứ điều gì nếu còn chưa suy nghĩ nghiêm túc về việc cá nhân tôi có thể làm gì để giải quyết điều đó, và tôi đã giữ lời hứa trong chuyến đi quảng bá sách tháng Một năm đó: Meia và tôi đã bàn bạc nhiều về việc lập ra một tổ chức phi lợi nhuận tập trung vào việc cải thiện tương lai sự sống thông qua quản lý công nghệ.

Cô ấy nhất định là chúng tôi phải đặt cho nó một cái tên khác xa với những thứ như “Viện Tận thế & U ám” hay “Viện Cùng-Lo-Lắng-Về-Tương-Lai-Nào”. Vì cái tên Viện Tương lai Nhân loại đã được sử dụng trước rồi, nên chúng tôi thống nhất đặt tên là Viện Tương lai Sự sống (FLI), với điểm cộng là bao gồm nhiều loài hơn. Ngày 22 tháng 1, chuyến công du quảng bá sách đưa chúng tôi tới Santa Cruz. Khi mặt trời California lặn xuống trên biển Thái Bình Dương cũng là lúc chúng tôi dùng bữa tối với người bạn già Anthony Aguirre và thuyết phục ông tham gia cùng. Ông không chỉ là một trong những người thông thái và sống theo lý tưởng nhất mà tôi biết, mà còn là người đã cùng tôi điều hành một tổ chức phi lợi nhuận khác, Viện Vấn đề Căn bản (xem trang <http://fqxi.org>), trong suốt hơn chục năm.

Tuần tiếp theo, chuyến đi đưa tôi đến London. Vì đang bị tương lai của AI choán lấy đầu óc, tôi đã liên lạc với Demis Hassabis, người tử tế mời tôi đến thăm trụ sở của DeepMind. Tôi choáng váng trước mức tăng trưởng của họ kể từ khi ông đến thăm tôi ở MIT hai năm trước đó. Google khi ấy vừa mua lại họ với giá khoảng 650 triệu đô-la Mỹ, và khi nhìn ngắm quang cảnh văn phòng rộng lớn của họ với đầy những bộ óc sáng láng theo đuổi mục tiêu lâu lĩnh của Demis là “giải mã trí tuệ”, tôi tin rằng thành công của FLI là một khả năng có thực.

Buổi tối sau đó, tôi trò chuyện với Jaan Tallinn qua Skype, phần mềm mà ông đã góp phần sáng tạo ra. Tôi giải thích về tầm nhìn của FLI và một giờ sau, ông quyết định đặt cửa thắng cho chúng tôi, rót vốn khoảng 100.000 đô-la mỗi năm! Ít có điều gì khiến tôi cảm động hơn việc có ai đó đặt vào tôi nhiều niềm tin hơn cả những gì tôi xứng đáng, nên một năm sau, sau hội nghị Puerto Rico mà tôi nhắc đến trong chương 1, khi ông nói đùa rằng đây là khoản đầu tư tốt nhất mà ông từng làm, tôi thấy như có cả thế giới bên mình vậy.

Hình 9.1: Hội nghị Asilomar tháng 1 năm 2017, phần tiếp theo của hội nghị Puerto Rico, đã hội tụ đông đảo các nhà nghiên cứu AI và các ngành liên quan. Hàng sau, từ trái sang phải:

Patrick Lin, Daniel Weld, Ariel Conn, Nancy Chang, Tom Mitchell, Ray Kurzweil, Daniel Dewey, Margaret Boden, Peter Norvig, Nick Hay, Moshe Vardi, Scott Siskind, Nick Bostrom, Francesca Rossi, Shane Legg, Manuela Veloso, David Marble, Katja Grace, Irakli Beridze, Marty Tenenbaum, Gill Pratt, Martin Rees, Joshua Greene, Matt Scherer, Angela Kane, Amara Angelica, Jeff Mohr, Mustafa Suleyman, Steve Omohundro, Kate Crawford, Vitalik Buterin, Yutaka Matsuo, Stefano Ermon, Michael Wellman, Bas Steunebrink, Wendell Wallach, Allan Dafoe, Toby Ord, Thomas Dietterich, Daniel Kahneman, Dario Amodei, Eric Drexler, Tomaso Poggio, Eric Schmidt, Pedro Ortega, David Leake, Seán Ó hÉigeartaigh, Owain Evans, Jaan Tallinn, Anca Dragan, Sean Legassick, Toby Walsh, Peter Asaro, Kay Firth-Butterfield, Philip Sabes, Paul Merolla, Bart Selman, Tucker Davey, ?, Jacob Steinhardt, Moshe Looks, Josh Tenenbaum, Tom Gruber, Andrew Ng, Kareem Ayoub, Craig Calhoun, Percy Liang, Helen Toner, David Chalmers, Richard Sutton, Claudia Passos-Ferreira, János Krámar, William MacAskill, Eliezer Yudkowsky, Brian Ziebart, Huw Price, Carl Shulman, Neil Lawrence, Richard Mallah, Jurgen Schmidhuber, Dileep George, Jonathan Rothberg, Noah Rothberg. Hàng trước: Anthony Aguirre, Sonia Sachs, Lucas Perry, Jeffrey Sachs, Vincent Conitzer, Steve Goose, Victoria Krakovna, Owen Cotton-Barratt, Daniela Rus, Dylan Hadfield-Menell, Verity Harding, Shivon Zilis, Laurent Orseau, Ramana Kumar, Nate Soares, Andrew McAfee, Jack Clark, Anna Salamon, Long Ouyang, Andrew Critch, Paul Christiano, Yoshua Bengio, David Sanford, Catherine Olsson, Jessica Taylor, Martina Kunz, Kristinn Thorisson, Stuart Armstrong, Yann LeCun, Alexander Tamas, Roman Yampolskiy, Marin Soljačić, Lawrence Krauss, Stuart Russell, Eric Brynjolfsson, Ryan Calo, ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, Max Tegmark, Adrian Weller, Jose Hernandez-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet,

Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt. Đến sau: Guru Banavar, Demis Hassabis, Rao Kambhampati, Elon Musk, Larry Page, Anthony Romero.

Ngày hôm sau, tôi dùng một thời gian trống trong lịch trình công du để đến thăm Bảo tàng Khoa học London. Sau một thời gian dài bị ám ảnh bởi quá khứ và tương lai của trí tuệ, tôi bỗng thấy mình như đang đi qua những hiện vật trưng bày chính những suy nghĩ của tôi. Họ đã khéo léo sắp xếp một bộ sưu tập kỳ thú tượng trưng cho sự phát triển của tri thức, gồm từ đầu máy xe lửa Rocket của Stephenson đến xe hơi Model T Ford, một bản sao kích thước như thật của tàu thám hiểm mặt trăng Apollo 11 và các loại máy tính từ thời máy tính cơ học “Máy Sai Phân” do Babbage chế tạo, đến các phần cứng ngày nay. Họ còn trưng bày lịch sử hiểu biết về trí óc, từ các thí nghiệm với chân ếch của Galvano đến nơ-ron, EEG và fMRI.

Tôi rất hiếm khi khóc, nhưng tôi đã rơi nước mắt khi bước ra khỏi bảo tàng – lại ngay giữa một đường hầm đầy người đi bộ trên đường ra nhà ga tàu điện ngầm South Kensington. Trước mắt tôi là những người đang hồ hởi đi theo nhịp sống thường ngày, hoàn toàn không biết tôi nghĩ gì trong đầu. Đầu tiên con người chúng ta phát hiện ra cách sao chép một vài quá trình tự nhiên bằng máy móc, tự tạo ra gió và tia chớp hay mã lực cơ khí của chính mình. Dần dần, chúng ta bắt đầu nhận ra cơ thể mình cũng là máy móc. Rồi việc phát hiện ra các tế bào thần kinh bắt đầu xóa mờ ranh giới giữa cơ thể và trí óc. Rồi chúng ta bắt đầu xây dựng các cỗ máy có thể vượt xa không chỉ cơ bắp của ta mà cả trí óc của ta nữa. Vậy song song với việc tìm ra chúng ta tạo nên từ cái gì, liệu chúng ta có đang không tránh khỏi việc khiến cho bản thân trở nên lạc hậu? Đó sẽ là một bi kịch.

Suy nghĩ này khiến tôi sợ hãi, nhưng nó cũng củng cố trong tôi quyết tâm giữ vững mục tiêu. Tôi cảm thấy chúng tôi cần thêm một người để hoàn thiện đội ngũ sáng lập FLI, người sẽ lãnh đạo một đội tình nguyện viên trẻ trung và đầy lý tưởng. Lựa chọn hợp lý là Viktoriya Krakovna, một sinh viên sáng giá tốt nghiệp Harvard, không những giành huy chương bạc kỳ thi Olympic Toán Quốc tế mà còn sáng lập ra Citadel, một ngôi nhà cho khoảng một tá bạn trẻ sống

theo lý tưởng và muốn có một lý do để đóng góp nhiều hơn cho cuộc đời và cho thế giới. Năm ngày sau, Meia và tôi mời cô gái ấy đến nhà để trao đổi về tầm nhìn của chúng tôi, và trước khi chúng tôi kịp ăn hết món sushi, FLI đã ra đời.

Phiêu lưu tới Puerto Rico

Sự kiện này đánh dấu sự khởi đầu của một chuyến phiêu lưu thú vị hiện vẫn đang tiếp diễn. Như tôi đã kể trong chương 1, vợ chồng tôi tổ chức các buổi hội ý thường kỳ tại nhà mình với hàng chục sinh viên và các vị giáo sư đầy lý tưởng cùng các nhà tư duy khác trong vùng. Tại những buổi hội ý này, các ý tưởng hàng đầu đã biến thành những dự án – dự án đầu tiên là bài báo trang đầu về AI ở chương 1 cùng với Stephen Hawking, Stuart Russell và Frank Wilczek, đã giúp châm ngòi cho các cuộc tranh luận trong cộng đồng. Song song với những bước chập chững để thành lập một tổ chức mới (như kết nối, chiêu mộ ban cố vấn và tạo lập trang web), chúng tôi cũng tổ chức một lễ ra mắt vui vẻ trước thính phòng chật ních tại MIT, nơi Alan Alda khám phá tương lai công nghệ cùng các chuyên gia hàng đầu.



Hình 9.2: Jaan Tallinn, Anthony Aguirre, tác giả, Meia Chita-Tegmark và Viktoriya Krakovna kỷ niệm việc thành lập FLI bằng bữa sushi vào ngày 23 tháng 5 năm 2014.

Chúng tôi tập trung toàn bộ những tháng còn lại của năm đó để tổ chức hội thảo Puerto Rico với mục tiêu, như tôi đề cập trong chương 1, là thu hút các nhà nghiên cứu AI hàng đầu trên thế giới vào cuộc thảo luận về cách giữ cho AI luôn hữu ích. Mục đích của chúng tôi là

đưa cuộc đối thoại về an toàn AI từ sự lo lắng sang hành động: từ tranh cãi vấp vãnh về các nỗi lo, trở thành thống nhất về những dự án nghiên cứu vững chắc có thể bắt đầu ngay lập tức để tối đa hóa cơ hội đem lại kết quả tốt. Để chuẩn bị, chúng tôi thu thập những ý tưởng nghiên cứu nhiều hứa hẹn về an toàn AI trên khắp thế giới và thu thập những lời nhận xét từ cộng đồng về danh sách dự án đang dần lớn lên của mình. Với sự giúp đỡ của Stuart Russell và một nhóm tình nguyện trẻ chăm chỉ, đặc biệt là Daniel Dewey, János Krámar và Richard Mallah, chúng tôi chốt lọc những nghiên cứu này thành tài liệu để thảo luận tại hội nghị.¹ Chúng tôi hy vọng xây dựng sự đồng thuận về việc có rất nhiều nghiên cứu an toàn AI mang giá trị và cần tiến hành sẽ khuyến khích mọi người bắt tay vào những nghiên cứu như thế. Mục tiêu cuối cùng là sẽ thuyết phục được ai đó tài trợ cho dự án này bởi, cho tới khi ấy, gần như không có hỗ trợ nào từ các cơ quan tài trợ thuộc chính phủ.

Rồi Elon Musk xuất hiện. Ngày 2 tháng 8, anh xuất hiện trong tầm ngắm của chúng tôi qua dòng tweet nổi tiếng: “Cuốn Siêu trí tuệ của Bostrom rất đáng đọc. Chúng ta cần vô cùng cẩn thận với AI. Có khả năng còn nguy hiểm hơn cả đồng hạt nhân.” Tôi liên hệ với anh, chia sẻ nỗ lực của chúng tôi và được trao đổi qua điện thoại vài tuần sau đó. Mặc dù tôi thấy khá hồi hộp và nôn nao, nhưng kết quả rất tuyệt: anh đồng ý tham gia vào ban cố vấn khoa học cho FLI, đồng ý tới dự hội nghị và có khả năng tài trợ cho chương trình nghiên cứu an toàn AI đầu tiên từ trước tới nay mà chúng tôi sẽ công bố tại Puerto Rico. Việc này khiến tất cả đội ngũ FLI vô cùng phấn khích và đẩy mạnh nỗ lực nhằm tạo ra một buổi hội nghị tuyệt vời, nhận định những chủ đề nghiên cứu hứa hẹn và xây dựng lực lượng ủng hộ cộng đồng.

Cuối cùng tôi cũng gặp mặt Elon trực tiếp để lên kế hoạch kỹ hơn, khi anh ấy tới MIT hai tháng sau đó để tham dự một hội nghị chuyên đề không gian. Cảm giác khá lạ lẫm khi ở một mình cùng Elon trong căn phòng nhỏ màu xanh lá, chỉ chốc lát sau khi anh vừa khiến hơn một nghìn sinh viên MIT bùng nổ như ngôi sao nhạc rock. Nhưng sau vài phút, trong đầu tôi chỉ còn nghĩ về dự án chung của cả hai. Tôi thích Elon ngay lập tức. Từ anh toả ra sự chân thành, và tôi thấy đầy cảm hứng trước sự quan tâm nhiệt thành của anh ấy về tương lai xa

của nhân loại, cũng như cách anh táo bạo biến khát vọng thành hành động. Vì muốn nhân loại khám phá và định cư trong Vũ trụ của ta mà Elon thành lập công ty không gian. Vì muốn có nguồn năng lượng bền vững mà Elon mở công ty năng lượng mặt trời và xe điện. Cao ráo, đẹp trai, ăn nói hùng hồn và cực kỳ hiểu biết, cũng dễ hiểu tại sao mọi người lắng nghe anh ấy.

Không may thay, sự kiện MIT này cũng dạy cho tôi biết rằng truyền thông dường như thích hướng vào nỗi sợ và gây bất hoà. Phần trình bày trên sân khấu của Elon là một cuộc thảo luận thú vị dài một tiếng về việc khám phá không gian, và tôi đã nghĩ sẽ là tin thời sự tuyệt vời. Vào phút cuối cùng, một sinh viên nêu ra câu hỏi ngoài lề về AI. Phần trả lời của Elon có một câu “với trí tuệ nhân tạo, chúng ta đang triệu hồi ác quỷ”, lại là thứ duy nhất phần lớn truyền thông đưa tin – và đa phần đều tách khỏi ngữ cảnh. Nó khiến tôi giật mình nhận ra rất nhiều phóng viên đang vô ý làm chính thứ ngược lại với điều chúng tôi cố gắng đạt được ở Puerto Rico. Trong khi chúng tôi muốn xây dựng sự đồng thuận trong cộng đồng qua việc nhấn mạnh những điểm chung thì truyền thông lại có động cơ nhấn mạnh sự khác biệt. Tin bài của họ càng gây tranh cãi thì chỉ số Nielsen và doanh thu quảng cáo của họ càng cao. Hơn thế, trong khi chúng tôi muốn giúp những người với đủ loại quan điểm khác nhau tụ họp, kết nối với nhau và hiểu nhau hơn, thì bằng việc đăng tin, trích dẫn những câu mang tính khiêu khích mà không kèm ngữ cảnh, tin truyền thông lại vô tình khiến những người thuộc các phái quan điểm khác nhau thất vọng về nhau và kích thích sự bất đồng ý kiến. Vì lẽ đó, chúng tôi quyết định cấm phóng viên tới cuộc gặp Puerto Rico và thi hành “Luật Chatham House”, ngăn cấm những người tham dự tiết lộ ai đã nói gì.⁶⁶

Hội nghị Puerto Rico cuối cùng cũng thành công, nhưng không hề dễ dàng. Khoảng thời gian ngắn còn lại chủ yếu đòi hỏi việc chuẩn bị kỹ lưỡng, ví dụ tôi phải gọi điện hoặc trao đổi qua skype với rất nhiều nhà nghiên cứu AI để tập hợp nên một nhóm người tham gia có tầm ảnh hưởng nhằm thu hút thêm những người tham dự khác, và cũng có cả những khoảnh khắc “thót tim” – chẳng hạn như lần tôi thức dậy vào 7 giờ sáng ngày 27 tháng 12 để gọi cho Elon ở Uruguay dù

đường truyền điện thoại khá tệ, rồi được bảo rằng “Tôi không nghĩ việc này có thể thành.” Anh lo ngại chương trình nghiên cứu an toàn AI có thể đem lại cảm giác an toàn ngộ nhận, tạo điều kiện cho các nhà nghiên cứu liều lĩnh tiếp tục làm liều trong khi vẫn rao giảng về tính an toàn. Nhưng sau đó, mặc dù âm thanh liên tục bị nhiễu, chúng tôi vẫn trao đổi kỹ lưỡng về những lợi ích to lớn của việc phổ biến chủ đề này và thu hút thêm nhiều nhà nghiên cứu AI làm việc về an toàn AI. Sau khi cuộc gọi bị đứt hẳn, Elon gửi cho tôi một trong những email mà tôi yêu thích nhất: “Mất sóng ở đầu bên này. Dù sao thì, tài liệu có vẻ ổn. Tôi rất vui được hỗ trợ cho nghiên cứu 5 triệu đô-la trong ba năm. Hay chúng ta nên tăng lên 10 triệu đô-la? ”

Bốn ngày sau, năm 2015 tới với khởi đầu tốt đẹp cho cả Meia và tôi khi chúng tôi có quãng thời gian thư giãn ngắn trước cuộc họp, nhảy múa mừng năm mới trên bãi biển Puerto Rico sáng bừng dưới ánh pháo hoa. Hội nghị cũng có một khởi đầu tuyệt vời: số đông đáng kể đồng thuận rằng cần có nhiều nghiên cứu về an toàn AI hơn, và dựa trên ý kiến từ những người tham gia hội nghị, tài liệu về những nghiên cứu ưu tiên mà chúng tôi nỗ lực biên soạn đã được cải thiện và hoàn thành. Chúng tôi chuyền tay nhau bức thư ngỏ nhắn nhủ về nghiên cứu an toàn mà tôi đã nhắc đến ở chương 1, và rất vui khi gần như tất cả mọi người đều ký tên lên nó.

Trong cuộc gặp “thần kỳ” với Meia và tôi trong phòng khách sạn, Elon trao cho chúng tôi kế hoạch chi tiết về chương trình tài trợ. Meia cảm động vì cách anh nói chuyện thực tế và thẳng thắn về đời tư, cũng như vì cách anh quan tâm tới chúng tôi. Elon hỏi chúng tôi gặp nhau như thế nào và rất thích câu chuyện đông dài của Meia. Ngày hôm sau, chúng tôi quay lại bài phỏng vấn anh về an toàn AI và lý do vì sao anh muốn ủng hộ, mọi thứ dường như đều tốt đẹp.²

Điểm nhấn của hội nghị là tuyên bố về việc tài trợ của Elon, được chúng tôi xếp lịch lúc 7 giờ tối ngày Chủ nhật, 4 tháng 1 năm 2015. Tôi căng thẳng tới mức cứ trở mình suốt đêm hôm trước. Và rồi, chỉ 15 phút trước khi sắp bước vào phần quan trọng đó thì chúng tôi vấp phải trở ngại bất ngờ: trợ lý của Elon gọi đến và nói có khả năng Elon không thực hiện được việc tuyên bố! Meia nói cô ấy chưa bao giờ thấy tôi căng thẳng hay thất vọng đến vậy. Elon cuối cùng cũng tới,

còn tôi thì đếm ngược từng giây đếm ngược tới lúc mở màn khi chúng tôi ngồi đó và bàn bạc. Anh ấy giải thích là bên SpaceX chỉ hai ngày nữa là phóng thử tên lửa mà họ hy vọng sẽ thành công lần đầu tiên trong việc hạ cánh tầng thứ nhất của tên lửa lên tàu thủy không người lái, và vì đây là một cột mốc quan trọng nên đội ngũ SpaceX không muốn truyền thông bị phân tán với các tin tức song song liên quan đến Elon. Anthony Aguirre, bình tĩnh và điềm đạm như mọi khi, chỉ ra rằng điều này nghĩa là không ai muốn truyền thông chú ý tới việc này, dù là Elon nói riêng hay cộng đồng AI nói chung. Chúng tôi đến trễ vài phút cho phần hội nghị mà tôi sẽ chủ trì, nhưng chúng tôi đã có một kế hoạch: sẽ không đề cập đến một khoản tiền nào để đảm bảo việc tuyên bố không đáng đưa tin, đồng thời tôi sẽ bắt buộc tất cả mọi người theo Luật Chatham House để giữ bí mật tuyên bố của Elon trước thế giới trong chín ngày nếu như tên lửa tới được trạm không gian, bất kể việc hạ cánh có thành công hay không; anh nói thậm chí sẽ cần nhiều thời gian hơn nếu tên lửa phát nổ trong khi phóng.

Thời gian còn lại đến thời khắc tuyên bố cuối cùng cũng đã về tới số 0. Các cử tọa về siêu trí tuệ mà tôi đang thảo luận cùng vẫn còn ngồi ngay bên cạnh tôi trên sân khấu: Eliezer Yudkowsky, Elon Musk, Nick Bostrom, Richard Mallah, Murray Shanahan, Bart Selman, Shane Legg và Vernor Vinge. Tiếng vỗ tay ngớt dần, nhưng các cử tọa vẫn ngồi yên, vì tôi đã bảo họ ở lại mà không giải thích vì sao. Sau này Meia kể lại với tôi rằng tim cô khi ấy như vọt lên đến tận trời và cô phải bám lấy tay Viktoriya Krakovna dưới mặt bàn để bình tĩnh lại. Tôi cười, biết rằng đây chính là khoảnh khắc mà vì nó chúng tôi đã làm việc, hy vọng và chờ đợi bao lâu nay.

Tôi phát biểu rằng tôi rất hạnh phúc vì hội nghị đạt được sự nhất trí cao là cần thêm các nghiên cứu để giữ cho AI có ích, và rằng có nhiều hướng nghiên cứu vững chắc mà ta có thể bắt đầu tiến hành ngay. Nhưng trong phiên họp này đã có nhiều ý kiến về những rủi ro nghiêm trọng, tôi nói thêm, nên nếu có thể nâng cao tinh thần và khởi động tâm trạng sôi nổi trước khi bước ra bữa tiệc đã bày sẵn ngoài kia thì thật là hay. “Và vì thế tôi xin nhường micro cho... Elon Musk!” Tôi cảm thấy như lịch sử đang thành hình khi Elon cầm lấy mic và tuyên bố anh sẽ ủng hộ một khoản tiền lớn cho các nghiên cứu về an

toàn AI. Không có gì ngạc nhiên khi anh khiến cả hội trường vỡ òa. Theo kế hoạch, anh không đề cập số tiền là bao nhiêu, nhưng tôi thì biết đó là con số 10 triệu đô-la tuyệt vời như chúng tôi đã thỏa thuận.

Meia và tôi đi thăm ông bà nội ngoại ở Thụy Điển và Rumania sau hội nghị, cùng cha tôi nín thở theo dõi màn phóng tên lửa được live stream ở Stockholm. Nỗ lực hạ cánh không may kết thúc bằng cái mà Elon nói giảm nói tránh là RUD, “tách rời nhanh không dự kiến”, và để hạ cánh xuống biển thành công, đội ngũ của ông phải mất thêm 15 tháng nữa.³ Tuy nhiên, tất cả vệ tinh đều đã được phóng thành công vào quỹ đạo, cũng như chương trình tài trợ của chúng tôi được Elon đăng lên Twitter cho hàng triệu người theo dõi anh cùng biết.⁴

Phổ biến an toàn AI

Một mục tiêu chủ chốt của hội nghị Puerto Rico là để đưa nghiên cứu an toàn AI thành trào lưu phổ biến, và nhìn thấy mục tiêu này được thực hiện theo từng bước đem lại niềm hào hứng thật lớn. Phải kể đến đầu tiên là bản thân cuộc hội nghị đã giúp nhiều nhà nghiên cứu bắt đầu cảm thấy thoải mái khi trò chuyện về chủ đề này, một khi họ nhận ra mình là một phần trong một cộng đồng ngày càng phát triển. Tôi cảm động sâu sắc trước sự khích lệ từ rất nhiều người tham dự. Ví dụ, giáo sư AI của Đại học Cornell, Bart Selman, đã gửi thư cho tôi nói rằng: “Tôi thật lòng chưa từng thấy một cuộc gặp gỡ khoa học nào được tổ chức tốt hơn, đem đến tinh thần phấn chấn và kích thích trí tuệ hơn thế.”

Bước tiếp theo bắt đầu vào ngày 11 tháng 1 khi Elon đăng tải trên Twitter: “Các nhà phát triển trí tuệ nhân tạo hàng đầu thế giới ký tên vào thư ngỏ kêu gọi nghiên cứu về an toàn AI,”⁵ kèm theo một liên kết đưa người đọc đến trang ký tên, chẳng mấy chốc đã thu về hơn tám nghìn chữ ký, bao gồm nhiều nhà chế tạo AI nổi bật nhất trên thế giới. Bỗng nhiên người ta thấy còn khó hơn để mà cho rằng những người quan tâm đến an toàn AI không biết họ đang nói về cái gì, vì giờ đây nói như thế cũng hàm ý rằng đến các nhà nghiên cứu AI hàng đầu cũng không biết họ đang nói về cái gì. Theo dõi cách bức

thư ngỏ được truyền thông khắp thế giới đưa tin, chúng tôi thầm cảm ơn vì đã cấm các nhà báo tham dự hội nghị. Mặc dù từ ngữ có tính cảnh báo nhất trong bức thư chỉ là “cạm bẫy”, nó vẫn gây ra những tiêu đề như “Elon Musk và Stephen Hawking ký tên vào thư ngỏ hy vọng ngăn chặn robot nổi dậy”, với hình ảnh minh họa là những cỗ máy hủy diệt đang sẵn sàng sát khí. Trong số hàng trăm bài báo đã bắt gặp, chúng tôi thích nhất một bài chế nhạo các bài báo khác, viết rằng “một tiêu đề khiến người ta hình dung các bộ xương máy móc giẫm đạp lên sọ người biến công nghệ phức tạp và mang tính cải cách trở thành một trò biểu diễn lễ hội đường phố.”⁶ May thay, cũng có rất nhiều bài báo đưa tin tinh táo và đem lại cho chúng tôi một thách thức mới là theo dõi cơn bão các chữ ký mới đổ về, vì cần phải xác minh các chữ ký này theo cách thủ công để bảo vệ mức độ đáng tin của chúng tôi và loại bỏ những cái tên đùa cợt như “HAL 9000”, “Kẻ hủy diệt”, “Sarah Jeanette Connor” và “Skynet.” Để làm việc này và chuẩn bị cho các bức thư ngỏ trong tương lai, Viktoriya Krakovna và János Krámar giúp tổ chức một đội tình nguyện kiểm tra gồm có Jesse Galef, Eric Gastfriend và Revathi Vinoth Kumar làm việc theo ca, để khi Revathi ở Ấn Độ đi ngủ thì sẽ có Eric ở Boston tiếp quản và cứ như vậy.

Mốc quan trọng thứ ba là bốn ngày sau đó, khi Elon đăng lên Twitter một liên kết tới tuyên bố rằng anh ủng hộ 10 triệu đô-la cho nghiên cứu an toàn AI.⁷ Một tuần sau, chúng tôi ra mắt cổng thông tin trực tuyến nơi các nhà nghiên cứu khắp thế giới có thể nộp đơn và tranh đua lấy vốn tài trợ. Chúng tôi có thể làm ra một hệ thống nộp đơn nhanh như vậy chỉ vì Anthony và tôi đã có cả chục năm điều hành các cuộc thi tương tự để xin tài trợ cho nghiên cứu vật lý. Dự án Thiện nguyện Mở, một tổ chức từ thiện có trụ sở ở California và tập trung vào hoạt động từ thiện có tác động lớn, đã hào phóng đồng ý hỗ trợ thêm vào món quà của Elon để chúng tôi có thể phân bổ nhiều nguồn tài trợ hơn. Chúng tôi không chắc sẽ nhận được bao nhiêu đơn đăng ký, vì chủ đề này còn rất mới lạ và hạn nộp thì gấp. Nhưng lượng phản hồi đã khiến chúng tôi choáng váng, với khoảng 300 đội nghiên cứu trên khắp thế giới xin được tài trợ tổng số tiền khoảng 100 triệu đô-la. Một ban giám khảo gồm các giáo sư AI và các

nhà nghiên cứu khác đã cẩn thận xem xét các đề xuất và chọn ra 37 đội thắng cuộc, được rót vốn nghiên cứu trong vòng ba năm. Khi chúng tôi công bố danh sách người thắng cuộc, đó là lần đầu tiên truyền thông đáp lại các hoạt động của chúng tôi bằng những tin tức mang màu sắc công bằng và không có hình ảnh minh họa kiểu robot sát thủ. Cuối cùng mọi người cũng bắt đầu hiểu được rằng an toàn AI không phải lời nói sáo rỗng: có những công việc thực sự có ích cần làm và nhiều đội nghiên cứu lớn đang xắn tay bắt đầu tham gia vào nỗ lực này. Bước quan trọng thứ tư diễn ra trong vòng hai năm tiếp theo, với hàng loạt bài nghiên cứu kỹ thuật được xuất bản và hàng chục lớp tập huấn về an toàn AI được tổ chức khắp thế giới, thường nằm trong các hội nghị AI chủ đạo. Những con người kiên định đã cố gắng nhiều năm để thu hút sự chú ý của cộng đồng AI vào nghiên cứu về an toàn và mới chỉ nhận được thành công hạn chế, nhưng giờ đây mọi thứ bắt đầu trên đà đi lên. Trong số các bài nghiên cứu được in có nhiều bài do chương trình tài trợ của chúng tôi cấp vốn. FLI cố gắng hết mình để tổ chức và cấp vốn cho càng nhiều lớp tập huấn càng tốt, nhưng tỷ lệ các lớp do các nhà nghiên cứu AI tự bỏ thời gian và tài chính tổ chức ngày càng tăng. Kết quả là, ngày càng có nhiều nhà nghiên cứu học hỏi thêm về nghiên cứu an toàn từ chính các đồng nghiệp của mình và phát hiện ra nghiên cứu này không chỉ có ích mà còn rất vui, liên quan đến các vấn đề tính toán và toán học thú vị để căng não giải đáp.

Dĩ nhiên không phải ai cũng thấy các phương trình phức tạp có gì vui. Hai năm sau hội nghị Puerto Rico, chúng tôi tổ chức một lớp tập huấn kỹ thuật trước thềm hội nghị Asilomar, trong đó những người giành được tài trợ FLI có thể trình bày nghiên cứu của mình. Tại đây, chúng tôi xem hết hình chiếu này đến hình chiếu khác với đầy những ký hiệu toán học trên màn hình lớn. Moshe Vardi, một giáo sư AI ở Đại học Rice, nói đùa rằng anh biết chúng tôi thành công trong việc lập ra ngành nghiên cứu an toàn AI ngay khi các cuộc họp bắt đầu trở nên buồn tẻ.

Sự tăng trưởng nhảy vọt của công trình nghiên cứu an toàn AI không giới hạn ở lĩnh vực học thuật. Amazon, DeepMind, Facebook, Google, IBM và Microsoft đã cho ra đời nhóm cộng tác trong ngành

để nghiên cứu AI có ích.⁸ Các khoản tài trợ mới và đáng kể cho an toàn AI đã giúp mở rộng nghiên cứu tại các tổ chức phi lợi nhuận “chị em” lớn nhất của chúng tôi: Viện nghiên cứu Trí tuệ máy ở Berkeley, Viện Tương lai Nhân loại ở Oxford và Trung tâm Nghiên cứu Rủi ro Sinh tồn ở Cambridge (Anh). Các khoản ủng hộ tiếp theo lên đến 10 triệu đô-la hay hơn, đã thúc đẩy thêm những nỗ lực vì AI có ích: Trung tâm Leverhulme vì Tương lai Trí tuệ ở Cambridge, Quỹ K&L Gates vì Đạo đức và Công nghệ tính toán ở Pittsburg, và Quỹ Đạo đức và Điều hành Trí tuệ Nhân tạo ở Miami được thành lập. Cuối cùng, với lời cam kết trị giá cả tỉ đô, Elon Musk bắt tay cùng các doanh nhân khác cho ra đời OpenAI, một công ty phi lợi nhuận ở San Francisco theo đuổi AI có ích. Nghiên cứu an toàn AI vậy là đã có chỗ đứng.

Song song chặt chẽ với việc công tác nghiên cứu tăng lên là sự nở rộ các quan điểm, ý kiến của cả các cá nhân và các đoàn thể được trình bày. Nhóm cộng tác ngành về AI đã cho in nguyên tắc nền tảng của nhóm, đồng thời chính phủ Mỹ, Đại học Stanford và IEEE (tổ chức lớn nhất thế giới của các chuyên gia kỹ thuật) cũng phát hành các báo cáo dài với danh sách những khuyến nghị, cùng hàng chục báo cáo và bài báo bày tỏ quan điểm từ các nơi khác.⁹

Chúng tôi mong ngóng được đẩy nhanh cuộc thảo luận đầy ý nghĩa giữa các đại biểu tham dự hội nghị Asilomar và biết được cộng đồng đa dạng này nhất trí về điều gì, nếu có. Lucas Perry do đó nhận nhiệm vụ anh hùng là đọc hết và tóm gọn tất cả các quan điểm có trong những văn bản mà chúng tôi tìm thấy. Qua nỗ lực miệt mài do Anthony Aguirre khởi xướng và kết thúc bằng một loạt cuộc họp đàm thoại kéo dài, đội ngũ FLI đã cố gắng nhóm các quan điểm tương tự lại với nhau và lược bỏ những sự dông dài thừa thãi để còn lại một danh sách duy nhất gồm các nguyên tắc súc tích và những quan điểm có ảnh hưởng lớn nhưng chưa xuất bản mà mới chỉ được trình bày trong các bài nói chuyện hay các hình thức khác ít chính thống hơn. Nhưng danh sách này vẫn chứa nhiều sự mơ hồ, mâu thuẫn trái ngược và nhiều chỗ có thể diễn giải theo các cách khác nhau, nên trong vòng một tháng trước hội nghị, chúng tôi chia sẻ với những

người tham dự và thu thập ý kiến đóng góp của họ để có được các nguyên tắc cải tiến hay mới mẻ, từ đó đưa ra một danh sách nguyên tắc được sửa đổi bổ sung đáng kể để sử dụng tại hội nghị.

Tại Asilomar, danh sách tiếp tục được cải thiện theo hai bước. Đầu tiên, các nhóm nhỏ thảo luận những nguyên tắc mà họ quan tâm nhất (hình 9.4), đưa ra các chỉnh lý chi tiết, những phản hồi, nguyên tắc mới và các phiên bản cạnh tranh với các nguyên tắc cũ. Cuối cùng, chúng tôi khảo sát mọi người tham dự để xác định mức độ ủng hộ cho từng phiên bản của mỗi nguyên tắc.



Hình 9.3: Các nhóm gồm những nhà tư duy lớn nghiên ngẫm về các nguyên tắc AI ở Asilomar.

Quá trình làm việc tập thể này vừa thấu đáo vừa vất kiệt sức lực của chúng tôi: Anthony, Meia và tôi cắt giảm giờ ngủ và giờ ăn trưa tại hội nghị để gấp rút tập hợp kịp thời mọi thứ cần thiết cho bước tiếp theo. Nhưng việc đó đồng thời cũng khiến chúng tôi vô cùng phấn chấn. Sau những cuộc thảo luận chi tiết, gai góc đôi khi đến độ nảy lửa và một loạt những phản hồi như vậy, chúng tôi ngỡ ngàng trước mức độ nhất trí cao đối với nhiều nguyên tắc trong lần khảo sát cuối cùng, trong đó một số nguyên tắc nhận được sự ủng hộ đến hơn 97%. Sự nhất trí này cho phép chúng tôi đặt tiêu chuẩn cao cho những nguyên tắc nào được đưa vào danh sách: chúng tôi chỉ giữ lại những nguyên tắc được ít nhất 90% người tham dự đồng ý. Mặc dù điều này nghĩa là một số nguyên tắc phổ biến bị bỏ đi vào phút cuối, kể cả một vài nguyên tắc mà cá nhân tôi thích,¹⁰ nhưng nó cũng

khiến phần lớn người tham dự cảm thấy thoải mái với việc ký tên tán thành tất cả các nguyên tắc trên tờ giấy đăng ký mà chúng tôi chuyền tay nhau quanh hội trường. Sau đây là kết quả.

Các nguyên tắc AI Asilomar

Trí tuệ nhân tạo đã cung cấp các công cụ hữu ích được nhiều người trên thế giới sử dụng hàng ngày. Việc tiếp tục phát triển nó, được định hướng bằng các nguyên tắc sau, sẽ đem lại cơ hội tuyệt vời để giúp đỡ và trao quyền cho con người trong các thập kỷ và thế kỷ sắp tới.

CÁC VẤN ĐỀ NGHIÊN CỨU

§1 Mục tiêu nghiên cứu: Mục tiêu của nghiên cứu AI nên là sáng tạo ra trí tuệ có ích chứ không phải là trí tuệ vô định hướng.

§2 Vốn nghiên cứu: Đầu tư vào nghiên cứu AI nên đi kèm với cấp vốn cho nghiên cứu về việc đảm bảo công dụng hữu ích của AI, bao gồm các vấn đề gai góc trong khoa học máy tính, kinh tế, pháp lý, đạo đức và xã hội học, như:

Làm thế nào để khiến hệ thống AI tương lai có độ vững chắc cao, sao cho chúng làm điều ta muốn mà không xảy ra lỗi hay bị thâm nhập phá hoại?

Làm thế nào để phát triển thịnh vượng nhờ tự động hóa trong khi vẫn duy trì nguồn tài nguyên và mục đích sống cho mọi người?

Làm thế nào để cập nhật hệ thống pháp lý cho công bằng và hiệu quả hơn, để từ đó theo kịp AI và quản lý được các rủi ro liên quan đến AI?

AI nên hòa hợp theo bộ giá trị nào, và nên có địa vị nào về mặt pháp lý và đạo đức?

§3 Liên kết giữa khoa học và chính sách: Nên có sự trao đổi lành mạnh và mang tính xây dựng giữa các nhà nghiên cứu AI và những người hoạch định chính sách.

§4 Văn hóa nghiên cứu: Một văn hóa hợp tác, tin tưởng và minh bạch nên được củng cố giữa các nhà nghiên cứu và phát triển AI.

§5 Tránh chạy đua: Các đội ngũ phát triển hệ thống AI nên chủ động hợp tác để tránh đi tắt trong các tiêu chuẩn về an toàn.

ĐẠO ĐỨC VÀ GIÁ TRỊ

§6 An toàn: Các hệ thống AI nên an toàn và bảo mật trong suốt quãng đời vận hành, đồng thời có thể xác thực độ an toàn và bảo mật trong những trường hợp kiểm chứng được và áp dụng được.

§7 Minh bạch về thất bại: Nếu một hệ thống AI gây hại, nên xác định chắc chắn nguyên nhân.

§8 Minh bạch về phân xử: Việc một hệ thống tự hành bất kỳ nào tham gia vào quyết định phân xử đều nên đi kèm một lời giải thích thỏa đáng mà một người có thẩm quyền có thể kiểm chứng được.

§9 Trách nhiệm: Các nhà thiết kế và chế tạo hệ thống AI trình độ cao đều có liên đới đến những hệ quả đạo đức của việc sử dụng, lạm dụng và các hành động của hệ thống AI đó, với trách nhiệm và cơ hội định hình những hệ quả đó.

§10 Hòa hợp giá trị: Các hệ thống AI tự hành cao nên được thiết kế sao cho mục tiêu và hành vi của chúng được đảm bảo là hòa hợp với các giá trị của con người trong suốt thời kỳ vận hành.

§11 Giá trị của con người: Các hệ thống AI nên được thiết kế và vận hành tương thích với các lý tưởng về phẩm giá, quyền, quyền tự do và sự đa dạng văn hóa của con người.

§12 Quyền riêng tư cá nhân: Mọi người nên có quyền tiếp cận, quản lý và kiểm soát dữ liệu họ tạo ra, nếu các hệ thống AI có năng lực phân tích và sử dụng các dữ liệu đó.

§13 Quyền tự do và bảo mật: Việc ứng dụng AI với dữ liệu cá nhân không được làm giảm sút sự tự do của mọi người, dù là trên thực tế hay theo cách nhìn nhận của họ.

§14 Lợi ích chung: Các công nghệ AI nên đem lại lợi ích cho càng nhiều người càng tốt.

§15 Tài sản chung: Sự thịnh vượng về kinh tế do AI tạo ra nên được chia sẻ rộng rãi, để đem lại lợi ích cho toàn thể nhân loại.

§16 Con người kiểm soát: Con người nên luôn có nên giao phó các quyết định cho các hệ thống AI để đạt được các mục tiêu do con người lựa chọn; nếu có, thì chọn cách nào.

§17 Không làm thoái hóa: Quyền lực có được nhờ việc kiểm soát các hệ thống AI tiên tiến trình độ cao nên tôn trọng và cải thiện, chứ

không làm thoái hóa, các quá trình xã hội và dân sự mà sức khỏe của xã hội phụ thuộc vào.

§18 Chạy đua vũ trang AI: Nên tránh chạy đua vũ trang các loại vũ khí tự hành chết người.

CÁC VẤN ĐỀ DÀI HẠN

§19 Thận trọng với năng lực: Không có sự nhất trí ở đây, vì vậy ta nên tránh những giả định chắc chắn về giới hạn trên của năng lực AI tương lai.

§20 Tầm quan trọng: AI tiên tiến có thể đại diện cho sự thay đổi sâu rộng trong lịch sử sự sống trên Trái Đất, và nên được quy hoạch và quản lý với sự quan tâm cùng nguồn tài nguyên tương xứng.

§21 Rủi ro: Rủi ro do các hệ thống AI đặt ra, đặc biệt là rủi ro mang tính thảm họa hay liên quan đến sự sinh tồn của con người, phải được đặt trong các nỗ lực quy hoạch và quản lý tương xứng với tác động dự kiến.

§22 Tự cải tiến đệ quy: Các hệ thống AI được thiết kế để có thể tự cải tiến hay tự tái tạo đệ quy theo cách có thể dẫn đến tăng nhanh chất lượng hay số lượng phải đi cùng các biện pháp an toàn và kiểm soát nghiêm ngặt.

§23 Tốt cho mọi người: Siêu trí tuệ chỉ nên được phát triển để phục vụ những lý tưởng đạo đức được chia sẻ rộng rãi, và vì lợi ích của toàn bộ nhân loại thay vì một nhà nước hay tổ chức.

Danh sách chữ ký tăng lên trông thấy sau khi chúng tôi đăng các nguyên tắc này lên mạng, và giờ nó đã khá đáng nể, bao gồm chữ ký của hơn 1.000 nhà nghiên cứu AI và nhiều nhà tư tưởng hàng đầu khác. Nếu bạn cũng muốn tham gia đóng góp chữ ký, bạn có thể ký tên tại: <http://futureoflife.org/ai-principles>.

Chúng tôi bất ngờ không chỉ bởi mức độ nhất trí về các nguyên tắc mà còn bởi mức độ mạnh mẽ của các nguyên tắc đó. Phải, thoát nhìn một số nghe có vẻ gây tranh cãi không kém gì câu “Hòa bình, tình yêu và việc làm mẹ đều giá trị”. Nhưng nhiều nguyên tắc thực sự sắc bén – bạn có thể thấy được dễ nhất bằng cách thử viết những câu phủ định chúng. Ví dụ, câu “Siêu trí tuệ là bất khả thi!” vi phạm điều §19, còn “Nghiên cứu về việc giảm rủi ro sinh tồn đến từ AI là hoàn toàn lãng phí!” vi phạm điều §21.

Quả thực, như bạn có thể tự mình thấy được nếu xem phần thảo luận dài hạn của chúng tôi trên YouTube,¹¹ Elon Musk, Stuart Russell, Ray Kurzweil, Demis Hassabis, Sam Harris, Nick Bostrom, David Chalmers, Bart Selman và Jaan Tallinn đều đồng ý rằng siêu trí tuệ có lẽ sẽ được phát triển và nghiên cứu về an toàn là quan trọng.

Lạc quan trong sự tỉnh táo

Như tôi đã thú thực ở đoạn đầu của phần kết này, tôi đang cảm thấy lạc quan về tương lai sự sống hơn tôi đã từng trong một thời gian dài. Tôi đã chia sẻ câu chuyện của cá nhân mình để giải thích tại sao.

Các trải nghiệm của tôi trong vài năm qua đã tăng niềm lạc quan cho tôi vì hai lý do riêng biệt. Đầu tiên, tôi đã chứng kiến cộng đồng AI đoàn kết lại đáng kể để cùng nhau đối mặt với những thách thức phía trước, thường là bằng cách hợp tác với các nhà tư duy ở các lĩnh vực khác. Elon bảo tôi sau cuộc gặp gỡ Asilomar rằng anh thấy thật kỳ diệu khi chỉ trong vài năm, an toàn AI đã đi từ một vấn đề ngoài lề thành trào lưu phổ biến, và tôi cũng thấy kỳ diệu y như vậy. Và giờ không chỉ những vấn đề ngắn hạn trong chương 3 trở thành các chủ đề thảo luận được coi trọng, mà cả siêu trí tuệ và rủi ro sinh tồn nữa, như được nêu trong Các nguyên tắc AI Asilomar. Không đời nào những nguyên tắc đó có thể được chấp thuận ở Puerto Rico chỉ hai năm về trước, khi từ nghe có vẻ đáng sợ nhất được đưa vào thư ngỏ là từ “cạm bẫy”.

Tôi thích quan sát mọi người và trong buổi sáng cuối cùng ở hội nghị Asilomar, tôi đã dành một lúc đứng bên rìa hội trường và nhìn ngắm các đại biểu tham dự đang lắng nghe một cuộc thảo luận về AI và luật pháp. Tôi ngạc nhiên khi thấy một cảm giác ấm áp dễ chịu lan tỏa trong người và bỗng nhiên thấy rất xúc động. Cảm giác này thật khác với Puerto Rico! Hồi đó, tôi nhớ mình nhìn phần lớn cộng đồng AI với cảm xúc tôn trọng và sợ hãi lẫn lộn – họ không hẳn là đối thủ, nhưng là một nhóm người mà tôi và các đồng nghiệp lo ngại về AI cảm thấy cần phải thuyết phục. Nhưng giờ cảm giác rõ ràng là chúng tôi đều đang ở trong cùng một đội. Như bạn có lẽ đã nhận ra đâu đó trong cuốn sách này, tôi vẫn chưa có các câu trả lời cho việc làm thế

nào để tạo ra một tương lai tươi sáng với AI, nên cảm giác thật tuyệt vời khi được là một phần trong một cộng đồng đang dần lớn mạnh và cùng nhau tìm kiếm các câu trả lời đó.



Hình 9.4: Một cộng đồng đang dần lớn mạnh cùng nhau tìm kiếm các câu trả lời tại Asilomar.

Lý do thứ hai khiến tôi trở nên lạc quan hơn là trải nghiệm FLI đã khiến tôi tự tin lên. Điều khiến tôi chảy nước mắt ở London là cảm giác về điều không thể tránh khỏi: rằng một tương lai u ám có thể đang tới và ta không thể làm gì được hết. Nhưng ba năm sau đó đã hóa giải nỗi u sầu trong tôi về định mệnh. Nếu chỉ một nhóm tình nguyện viên rách rưới cũng có thể tạo nên khác biệt tích cực đối với cái có thể nói là cuộc đối thoại quan trọng nhất của thời đại chúng ta, thì thử tưởng tượng xem những gì ta có thể làm khi ta hợp tác cùng nhau!

Erik Brynjolfsson nói về hai loại lạc quan trong bài phát biểu ở Asilomar. Đầu tiên là loại vô điều kiện, như mong đợi tích cực là sáng mai Mặt Trời sẽ lại mọc. Còn lại là loại mà ông gọi là “lạc quan trong sự tỉnh táo”, nghĩa là mong đợi rằng điều tốt đẹp sẽ xảy đến nếu bạn lên kế hoạch cẩn thận và làm việc chăm chỉ để đạt được điều đó. Đó là loại lạc quan mà tôi đang cảm thấy về tương lai sự sống.

Vậy bạn có thể làm gì để tạo nên khác biệt tích cực cho tương lai sự sống khi ta bước vào kỷ nguyên AI? Vì những lý do mà tôi sẽ giải thích ngay sau đây, tôi nghĩ rằng bước quan trọng đầu tiên là nỗ lực trở thành người lạc quan tỉnh táo, nếu bây giờ bạn chưa có niềm lạc quan đó. Để làm được điều này, điều quan trọng là phát triển tầm

nhìn tích cực về tương lai. Khi các sinh viên MIT tìm đến tôi ở văn phòng để xin tư vấn về sự nghiệp, tôi thường bắt đầu bằng cách hỏi họ rằng họ thấy mình ở đâu sau mười năm nữa. Nếu câu trả lời là “Có lẽ em sẽ nằm trong phòng điều trị ung thư, hay trong nghĩa trang sau khi bị xe buýt đâm”, tôi sẽ lên lớp sinh viên đó ra trò. Hình dung ra toàn những tương lai tiêu cực là một cách tiếp cận quá tồi khi lập kế hoạch cho sự nghiệp! Dành toàn bộ 100% nỗ lực của bản thân để né tránh bệnh tật và tai nạn là công thức hoàn hảo cho các chứng bệnh hoang tưởng, chứ không phải cho hạnh phúc. Thay vào đó, tôi muốn nghe sinh viên mô tả các mục tiêu với lòng nhiệt tình, sau đó chúng tôi có thể thảo luận các chiến lược để đi tới đó đồng thời tránh những cạm bẫy trên đường đi.

Erik đã chỉ ra rằng theo lý thuyết trò chơi, các hình dung tích cực tạo nền tảng cho một tỷ lệ lớn những sự hợp tác có trên đời, từ hôn nhân hay sáp nhập công ty cho đến quyết định của các bang độc lập tạo thành Liên bang Hoa Kỳ. Suy cho cùng, tại sao lại phải hy sinh cái gì đó bạn có nếu bạn không tưởng tượng ra là mình sẽ có được nhiều hơn từ sự hợp tác đó? Điều này nghĩa là chúng ta nên tưởng tượng ra những tương lai tích cực không chỉ cho bản thân mình, mà cho cả xã hội và cho nhân loại. Nói cách khác, chúng ta cần thêm hy vọng sinh tồn! Vậy mà, như Meia vẫn hay nhắc nhở tôi, từ Frankenstein đến Kẻ hủy diệt, tầm nhìn về tương lai trong văn học và điện ảnh hầu hết đều là phản không tưởng. Nói cách khác, xã hội chúng ta đang lập kế hoạch cho tương lai cũng tệ như sinh viên MIT mà tôi ví dụ ở trên. Đó là lý do chúng ta cần nhiều người lạc quan trong tình tảo hôn. Và đó là lý do trong suốt cuốn sách này tôi đã khuyến khích bạn nghĩ về tương lai mà bạn muốn thay vì chỉ là tương lai mà bạn sợ, để chúng ta có thể tìm ra các mục tiêu chung, từ đó lên kế hoạch và làm việc vì những mục tiêu đó.

Ta đã thấy trong cả cuốn sách này rằng AI sẽ có khả năng đem lại cho chúng ta những cơ hội lớn và thách thức khó nhằn như thế nào. Một chiến lược có khả năng giúp ích cho gần như mọi thách thức về AI là chúng ta bắt tay vào việc và cải thiện xã hội loài người trước khi AI lấy đà cất cánh trọn vẹn. Sẽ tốt hơn nếu chúng ta giáo dục lớp trẻ phải làm cho công nghệ vững chắc và có ích trước khi giao phó quyền lực lớn cho nó. Sẽ tốt hơn nếu chúng ta hiện đại hóa luật pháp

trước khi công nghệ khiến luật pháp lạc hậu. Sẽ tốt hơn nếu chúng ta giải quyết xung đột quốc tế trước khi chúng leo thang thành cuộc chạy đua vũ trang vũ khí tự hành. Sẽ tốt hơn nếu chúng ta tạo ra nền kinh tế đảm bảo thịnh vượng cho tất cả trước khi AI có tiềm năng làm sâu thêm sự bất bình đẳng. Sẽ tốt hơn nếu xã hội chúng ta thực thi chứ không ngó lơ các kết quả nghiên cứu an toàn AI. Và nhìn xa hơn về phía trước, tới những thách thức liên quan đến AGI siêu phàm, sẽ tốt hơn nếu chúng ta đồng thuận về ít nhất là một số tiêu chuẩn đạo đức cơ bản trước khi bắt đầu dạy dỗ những tiêu chuẩn này cho các cỗ máy cực mạnh. Trong một thế giới phân cực và hỗn loạn, những người có quyền lực sử dụng AI cho mục đích xấu xa sẽ có nhiều động lực và khả năng để làm vậy hơn, và các đội chạy đua để chế tạo AGI sẽ cảm thấy đi tắt đón đầu trong việc đảm bảo an toàn thì áp lực hơn là hợp tác với nhau. Tóm lại, nếu ta có thể tạo ra xã hội loài người hòa thuận hơn, nơi con người cùng hợp tác đi đến các mục tiêu chung, thì tiềm năng kết thúc tốt đẹp của cuộc cách mạng AI sẽ được cải thiện.

Nói cách khác, một trong những cách tốt nhất để bạn cải thiện tương lai của sự sống chính là cải thiện ngày mai. Bạn có quyền lực để làm vậy theo nhiều cách khác nhau. Dĩ nhiên bạn có thể bỏ phiếu trong hòm bầu cử và nói cho các chính trị gia biết bạn nghĩ gì về giáo dục, quyền riêng tư, vũ khí tự hành chết người, thất nghiệp do công nghệ và các vấn đề khác. Nhưng bạn cũng bỏ phiếu mỗi ngày, thông qua những gì bạn chọn mua, tin tức bạn chọn nghe, những gì bạn chọn chia sẻ và mẫu người bạn chọn để trở thành. Bạn muốn là người xen vào tất cả các cuộc đối thoại bằng cách xem điện thoại, hay là người cảm thấy được trao quyền khi sử dụng công nghệ chủ động và có kế hoạch? Bạn muốn sở hữu công nghệ hay muốn công nghệ sở hữu bạn? Bạn muốn việc làm người có ý nghĩa gì trong kỷ nguyên của AI? Xin hãy thảo luận những điều này với những người xung quanh bạn – đó là một cuộc đối thoại không chỉ quan trọng mà còn cực kỳ thú vị.

Chúng ta là những vệ binh của tương lai sự sống ngay từ bây giờ, khi ta định hình cho kỷ nguyên AI. Dù đã khóc ở London, nhưng giờ đây tôi cảm thấy không có gì là không thể tránh khỏi về tương lai này cả, và tôi biết việc tạo nên sự khác biệt dễ hơn tôi tưởng nhiều.

Tương lai của chúng ta không được tạc vào đá để chúng ta chỉ việc chờ đợi là sẽ xảy ra – tương lai do chúng ta tạo ra. Hãy cùng tạo ra một tương lai đầy cảm hứng!

www.facebook.com/groups/yeukindlevietnam



I LOVE
READING
BOOKS

SO MANY
BOOKS
SO LITTLE
TIME

the SWEET
SERENITY
of BOOKS



GIVE
me
a BOOK

A PAGE
A DAY



Chú thích

Chương 1

[←1]

“The AI Revolution: Our Immortality or Extinction?” theo trang Wait But Why (27 tháng 1, 2015), tại <http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html>.

[←2]

Bức thư ngỏ này, “Research Priorities for Robust and Beneficial Artificial Intelligence,” có thể tìm thấy tại <http://futureoflife.org/ai-open-letter/>.

[←3]

Ví dụ cho việc truyền thông gieo hoang mang về rô-bốt cổ điển: Ellie Zolfagharifard, “Artificial Intelligence ‘Could Be the Worst Thing to Happen to Humanity,’ ” báo Daily Mail, 2 tháng 5, 2014; <http://tinyurl.com/hawkingbots>.

Chương 2

Chú giải về nguồn gốc của thuật ngữ AGI:

<http://wp.goertzel.org/who-coined-the-term-agi>.

Hans Moravec, “When Will Computer Hardware Match the Human Brain?” tạp chí Journal of Evolution and Technology (1998), quyển 1.

Trong biểu đồ thể hiện năng lực tính toán theo năm, dữ liệu trước năm 2011 lấy từ quyển How to Create a Mind của Ray Kurzweil, và các dữ liệu sau đó được tính toán dựa trên nguồn thông tin tại <https://en.wikipedia.org/wiki/FLOPS>.

Nhà tiên phong về điện toán lượng tử David Deutsch nhìn nhận tính toán lượng tử như bằng chứng về vũ trụ song song trong cuốn The Fabric of Reality: The Science of Parallel Universes—

and Its Implications (London: Allen Lane, 1997). Nếu bạn muốn biết cái nhìn của tôi về vũ trụ song song lượng tử như cấp thứ ba trong bốn cấp độ đa vũ trụ, bạn sẽ thấy trong cuốn sách trước của tôi Max Tegmark, *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality* (New York: Knopf, 2014).

Chương 3

Xem “Google DeepMind’s Deep Q-learning Playing Atari Breakout” trên YouTube tại <https://tinyurl.com/atariiai>.

Theo Volodymyr Mnih et al., “Human-Level Control Through Deep Reinforcement Learning,” *Nature* 518 (26 tháng 2, 2015): 529–533, bản trực tuyến tại <http://tinyurl.com/ataripaper>.

Đây là một video về rô-bốt Big Dog đang hoạt động: <https://www.youtube.com/watch?v=W1czBcnX1Ww>.

Về phản ứng trước nước cờ sáng tạo ấn tượng ở hàng thứ 5 của AlphaGO, hãy xem “Move 37!! Lee Sedol vs AlphaGo Match 2,” tại <https://www.youtube.com/watch?v=JNrXgpSEEIE>.

Demis Hassabis mô tả phản ứng trước AlphaGo của các kỳ thủ cờ vây con người: <https://www.youtube.com/watch?v=otJKzpNWZT4>.

Về các cải tiến mới trong lĩnh vực dịch máy, xem Gideon Lewis-Kraus, “The Great A.I. Awakening,” *New York Times Magazine* (14 tháng 12, 2016), bản trực tuyến tại <http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html>. Truy cập GoogleTranslate tại <https://translate.google.com>.

Cuộc thi Winograd Schema Challenge: <http://tinyurl.com/winogradchallenge>.

Video về vụ nổ Ariane 5: <https://www.youtube.com/watch?v=qnHn8W1Em6E>.

Báo cáo về lỗi Chuyển bay 501 Ariane 5 của ban điều tra: <http://tinyurl.com/arianeflop>.

Báo cáo của ban điều tra về tai nạn tàu NASA Mars Climate Orbiter giai đoạn I: <http://tinyurl.com/marsflop>.

Giải thích chi tiết và nhất quán nhất về nguyên nhân nhiệm vụ tới Sao Kim của Mariner 1 thất bại là do bản chép tay sai của một biểu tượng toán học duy nhất (thiếu dấu gạch đầu):

<http://tinyurl.com/marinerflop>.

Mô tả chi tiết về nhiệm vụ thất bại của tàu Phobos 1 Xô Viết có thể tìm thấy tại Wesley T. Huntress Jr. và Mikhail Ya. Marov, Soviet Robots in the Solar System (New York: Praxis Publishing, 2011), trang 308.

Phần mềm chưa kiểm định khiến hãng Knight Capital mất 440 triệu đô-la trong vòng 45 phút: <http://tinyurl.com/knightflop1> và <http://tinyurl.com/knightflop2>.

Báo cáo của chính phủ Hoa Kỳ về vụ “sụp đổ chóng vánh” ở phố Wall: “Findings Regarding the Market Events of May 6, 2010” (30 tháng 9, 2010), tại <http://tinyurl.com/flashcrashreport>.

In 3-D các tòa nhà (<https://www.youtube.com/watch?v=SObzNdyRTBs>), thiết bị vi cơ (<http://tinyurl.com/tinyprinter>) và những thứ có kích thước ở giữa (<https://www.youtube.com/watch?v=xVU4FLrsPXs>).

Bản đồ các phòng thí nghiệm chế tạo cộng đồng trên thế giới: <https://www.fablabs.io/labs/map>.

Bài báo về việc Robert Williams bị sát hại bởi rô-bốt công nghiệp: <http://tinyurl.com/williamsaccident>.

Bài báo về việc Kenji Urada bị sát hại bởi rô-bốt công nghiệp: <http://tinyurl.com/uradaaccident>.

Bài báo về việc công nhân của Volkswagen bị sát hại bởi rô-bốt công nghiệp: <http://tinyurl.com/baunatalaccident>.

Báo cáo của chính phủ Hoa Kỳ về tai nạn tử vong của công nhân: https://www.osha.gov/dep/fatcat/dep_fatcat.html.

Thống kê tử vong do tai nạn xe cộ: <http://tinyurl.com/roadsafety2> và <http://tinyurl.com/roadsafety3>.

Về vụ tử vong đầu tiên do xe tự lái Tesla, tìm đọc Andrew Buncombe, “Tesla Crash: Driver Who Died While on Autopilot Mode ‘Was Watching Harry Potter,’” tờ Independent (1 tháng 7, 2016), <http://tinyurl.com/teslacrashstory>. Tìm đọc báo cáo của Văn phòng Điều tra Sai sót của Cục quản lý An toàn Giao thông Đường bộ Quốc gia Hoa Kỳ <http://tinyurl.com/teslacrashreport>.

Tìm hiểu thêm về thảm hoạ phà Herald of Free Enterprise, tìm đọc R.B. Whittingham, The Blame Machine: Why Human Error Causes Accidents (Oxford, UK: Elsevier, 2004).

Phim tài liệu về vụ rơi máy bay Air France 447:

<https://www.youtube.com/watch?v=dpPkp8OGQFI>; báo cáo tai

nạn: <http://tinyurl.com/af447report>; phân tích khách quan:

<http://tinyurl.com/thomsonarticle>.

Báo cáo chính thức về vụ mất điện ở Hoa Kỳ - Canada năm 2003: <http://tinyurl.com/uscanadablackout>.

Báo cáo cuối cùng của Ủy ban Tổng thống về vụ tai nạn trên đảo Three Mile: <http://www.threemileisland.org/downloads/188.pdf>.

Nghiên cứu của Hà Lan cho thấy AI ngang ngửa bác sĩ X-quang con người trong chẩn đoán dựa trên MRI về ung thư tiền liệt tuyến: <http://tinyurl.com/prostate-ai>.

Nghiên cứu của Stanford cho thấy AI có thể vượt qua các nhà bệnh lý học con người trong chẩn đoán ung thư phổi:

<http://tinyurl.com/lungcancer-ai>.

Điều tra về tai nạn xạ trị bằng Therac-25:

<http://tinyurl.com/theracfailure>.

Báo cáo về vụ quá liều bức xạ gây chết người ở Panama do giao diện người dùng khó hiểu:

<http://tinyurl.com/cobalt60accident>.

Nghiên cứu về tác dụng phụ trong phẫu thuật bằng rô-bốt:

<https://arxiv.org/abs/1507.03518>.

Bài báo về số ca tử vong do dịch vụ chăm sóc y tế kém:

<http://tinyurl.com/medaccidents>.

Yahoo tạo ra tiêu chuẩn mới cho “vụ thâm nhập lớn” khi tuyên bố một tỷ (!) tài khoản người dùng đã bị xâm phạm:

<https://www.wired.com/2016/12/yahoo-hack-billion-users/>.

Bài báo trên New York Times về bản tuyên án tha bổng và bản sau đó về vụ sát hại KKK: <http://tinyurl.com/kkkacquittal>.

Nghiên cứu của Danziger et al. 2011

(<http://www.pnas.org/content/108/17/6889.full>), chứng minh các thẩm phán đang đối thì khắt khe hơn, bị Keren Weinshall-

Margela và John Shapard chỉ trích là còn thiếu sót

(<http://www.pnas.org/content/108/42/E833.full>), nhưng nhóm

Danziger nhấn mạnh tuyên bố của mình có cơ sở (<http://www.pnas.org/content/108/42/E834.full>).

Báo cáo của Pro Publica về phân biệt chủng tộc trong phần mềm dự đoán tái phạm tội: <http://tinyurl.com/robojudge>.

Sử dụng fMRI và các kỹ thuật quét não khác như bằng chứng tại toà vẫn còn gây tranh cãi, cũng như tính đáng tin cậy của những kỹ thuật này, mặc dù rất nhiều nhóm tuyên bố mức chính xác đạt trên 90%:

<http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00709/full>.

PBS làm bộ phim The Man Who Saved the World về sự kiện Vasili Arkhipov một mình ngăn chặn cuộc tấn công hạt nhân của Xô Viết: <https://www.youtube.com/watch?v=4VPY2SgyG5w>.

Câu chuyện Stanislav Petrov bỏ qua cảnh báo về việc Hoa Kỳ tấn công hạt nhân, coi đó là báo động giả đã được chuyển thể thành phim The Man Who Saved the World (đừng nhầm nó với bộ phim cùng tên tại chú thích trên), và Petrov đã được Liên Hiệp Quốc vinh danh và trao Giải thưởng Công dân Thế giới: <https://www.youtube.com/watch?v=IncSjwWQHMo>.

Thư ngỏ của các nhà nghiên cứu AI và rô-bốt về vũ khí tự vận hành: <http://futureoflife.org/open-letter-autonomous-weapons/>.

Một nhân viên chính phủ Hoa Kỳ có vẻ mong đợi cuộc chạy đua vũ trang AI: <http://tinyurl.com/workquote>.

Nghiên cứu về bất bình đẳng tài sản tại Hoa Kỳ từ năm 1931: <http://gabriel-zucman.eu/files/SaezZucman2015.pdf>.

Báo cáo của Oxfam về bất bình đẳng tài sản toàn cầu: <http://tinyurl.com/oxfam2017>.

Để tìm hiểu sâu hơn về giả thuyết công nghệ thúc đẩy bất bình đẳng, tìm đọc Erik Brynjolfsson và Andrew McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies (New York: Norton, 2014).

Bài báo trên tờ The Atlantic về việc mức lương của những người ít học sụt giảm: <http://tinyurl.com/wagedrop>.

Dữ liệu trên hình được lấy từ Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, Emmanuel Saez và Gabriel Zucman, The World Wealth and Income Database (<http://www.wid.world>), bao gồm cả tăng vốn.

Bài thuyết trình của James Manyika cho thấy thu nhập chuyển từ lao động sang vốn:

http://futureoflife.org/data/PDF/james_manyika.pdf.

Dự báo về công việc tự động hoá trong tương lai của Đại học Oxford (<http://tinyurl.com/automationoxford>) và McKinsey (<http://tinyurl.com/automationmckinsey>).

Video về đầu bếp rô-bốt: <https://www.youtube.com/watch?v=fE6i2OO6Y6s>.

Marin Soljai khám phá những lựa chọn này tại hội thảo năm 2016 Computers Gone Wild: Impact and Implications of Developments in Artificial Intelligence on Society:

<http://futureoflife.org/2016/05/06/computers-gone-wild/>.

Andrew McAfee đề xuất cách tạo ra nhiều công việc tốt hơn:

http://futureoflife.org/data/PDF/andrew_mcafee.pdf.

Bên cạnh rất nhiều bài báo học thuật tranh luận rằng “lần này sẽ khác” đối với những người thất nghiệp do công nghệ, video “Humans Need Not Apply” cũng nêu quan điểm tương tự theo cách ngắn gọn: <https://www.youtube.com/watch?v=7Pq-S557XQU>.

Cục Thống kê Lao động Hoa Kỳ:

<http://www.bls.gov/cps/cpsaat11.htm>.

Lập luận “lần này sẽ khác” đối với những người thất nghiệp do công nghệ: Federico Pistono, Robots Will Steal Your Job, but That’s OK (2012), <http://robotswillstealyourjob.com>.

Thay đổi về số lượng ngựa ở Hoa Kỳ:

<http://tinyurl.com/horsedecline>.

Phân tích tổng hợp cho thấy tỷ lệ thất nghiệp ảnh hưởng đến hạnh phúc: Maïke Luhmann et al., “Subjective Well-Being and Adaptation to Life Events: A Meta-Analysis,” Journal of Personality and Social Psychology 102, no. 3 (2012): 592; truy cập trực tuyến tại

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289759>.

Nghiên cứu về yếu tố làm tăng cảm giác hạnh phúc của con người: Angela Duckworth, Tracy Steen và Martin Seligman, “Positive Psychology in Clinical Practice,” Annual Review of Clinical Psychology 1 (2005): 629–651, tại <http://tinyurl.com/well>

beingduckworth. Weiting Ng và Ed Diener, “What Matters to the Rich and the Poor? Subjective Well-Being, Financial Satisfaction, and Postmaterialist Needs Across the World,” *Journal of Personality and Social Psychology* 107, no. 2 (2014): 326, tại <http://psycnet.apa.org/journals/psp/107/2/326>. Kirsten Weir, “More than Job Satisfaction,” *Monitor on Psychology* 44, no. 11 (December 2013), tại <http://www.apa.org/monitor/2013/12/job-satisfaction.aspx>.

Nhân khoảng 10¹¹ tế bào thần kinh, khoảng 10⁴ kết nối trên mỗi nơ-ron và khoảng một (100) xung điện mỗi nơ-ron trong một giây với nhau có thể gợi ý con số khoảng 10¹⁵ FLOPS (1 petaFLOPS) đủ để mô phỏng bộ não con người, nhưng nhiều biến còn chưa được hiểu rõ, bao gồm cả chi tiết thời gian phát xung điện và câu hỏi liệu các phần nhỏ của nơ-ron và xi-nap có cần cũng phải được mô phỏng không. Nhà khoa học máy tính của IBM Dharmendra Modha đã ước tính rằng 38 petaFLOPS là bắt buộc (<http://tinyurl.com/javln43>), trong khi nhà thần kinh học Henry Markram đã ước tính một người cần khoảng 1.000 petaFLOPS (<http://tinyurl.com/6rpohqv>). Nhà nghiên cứu AI Katja Grace và Paul Christiano lập luận rằng khía cạnh tốn kém nhất trong mô phỏng não không phải là tính toán mà là giao tiếp, và đó cũng là một nhiệm vụ nằm trong phạm vi mà những siêu máy tính tốt nhất hiện tại có thể thực hiện: <http://aiimpacts.org/about>.

Một ước tính thú vị về sức mạnh tính toán của bộ não con người: Hans Moravec “When Will Computer Hardware Match the Human Brain?” *Journal of Evolution and Technology*, vol. 1 (1998).

Chương 4

Xem video về chim máy đầu tiên Markus Fischer, “A Robot That Flies like a Bird,” TED Talk, tháng 7, 2011, tại https://www.ted.com/talks/a_robot_that_flies_like_a_bird.

Chương 5

Ray Kurzweil, *The Singularity Is Near* (Điểm kỳ dị gần kề) (New York: Viking Press, 2005).

Viễn cảnh “AI bảo mẫu” của Ben Goertzel được mô tả tại đây:

https://wiki.lesswrong.com/wiki/Nanny_AI.

Cuộc thảo luận về mối quan hệ giữa máy móc và con người, và liệu máy móc có phải là nô lệ của chúng ta, hãy xem Benjamin Wallace-Wells, “Boyhood,” *New York magazine* (20 tháng 5, 2015), bản trực tuyến tại <http://tinyurl.com/aislaves>.

Tội ác tinh thần được Nick Bostrom bàn luận trong cuốn *Superintelligence* (Siêu trí tuệ) và trong bài báo gần đây với nhiều chi tiết kỹ thuật hơn: Nick Bostrom, Allan Dafoe và Carrick Flynn, “Policy Desiderata in the Development of Machine Superintelligence” (2016),

<http://www.nickbostrom.com/papers/aipolicy.pdf>.

Matthew Schofield, “Memories of Stasi Color Germans’ View of U.S. Surveillance Programs,” *McClatchy DC Bureau* (26 tháng 6, 2013), trực tuyến tại <http://www.mcclatchydc.com/news/nation-world/national/article24750439.html>.

Về những chiêm nghiệm kích thích tư duy trước cách con người có thể được khuyến khích tạo ra kết quả mà không ai muốn, tôi khuyên bạn nên đọc “Meditations on Moloch,”

<http://slatestarcodex.com/2014/07/30/meditations-on-moloch>.

Biểu đồ thời gian tương tác của các lần chiến tranh hạt nhân suýt soát được phát động do tai nạn, xem Future of Life Institute, “Accidental Nuclear War: A Timeline of Close Calls,” tại <http://tinyurl.com/nukeoops>.

Về các khoản thanh toán bồi thường cho các nạn nhân thử nghiệm hạt nhân ở Hoa Kỳ, xem trang web của Bộ Tư pháp Hoa Kỳ, “Awards to Date 4/24/2015,” tại

<https://www.justice.gov/civil/awards-date-04242015>.

Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack, tháng 4, 2008, trực tuyến tại http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf.

Nghiên cứu độc lập bởi các nhà khoa học ở cả Hoa Kỳ và Xô viết cảnh báo Reagan và Gorbachev về nguy cơ mùa đông hạt

nhân: P. J. Crutzen và J. W. Birks, “The Atmosphere After a Nuclear War: Twilight at Noon,” *Ambio* 11, no. 2/3 (1982): 114–125. R. P. Turco, O. B. Toon, T. P. Ackerman, J. B. Pollack và C. Sagan, “Nuclear Winter: Global Consequences of Multiple Nuclear Explosions,” *Science* 222 (1983): 1283–1292. V. V. Aleksandrov và G. L. Stenchikov, “On the Modeling of the Climatic Consequences of the Nuclear War,” *Proceeding on Applied Mathematics* (Moscow: Computing Centre of the USSR Academy of Sciences, 1983), 21. A. Robock, “Snow and Ice Feedbacks Prolong Effects of Nuclear Winter,” *Nature* 310 (1984): 667–670.

Tính toán hiệu ứng khí hậu của chiến tranh hạt nhân toàn cầu: A. Robock, L. Oman and L. Stenchikov, “Nuclear Winter Revisited with a Modern Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences,” *Journal of Geophysical Research* 12 (2007): D13107.

Chương 6

Tìm hiểu thêm, hãy đọc Anders Sandberg, “Dyson Sphere FAQ,” tại [http:// www.aleph.se/nada/dysonFAQ.html](http://www.aleph.se/nada/dysonFAQ.html).

Bài báo có tầm ảnh hưởng của Freeman Dyson về quả cầu mang tên ông: Freeman Dyson, “Search for Artificial Stellar Sources of Infrared Radiation,” *Science*, vol. 131 (1959): 1667–1668.

Louis Crane và Shawn Westmoreland giả thích về đề xuất động cơ lỗ đen trong “Are Black Hole Starships Possible?,” tại <http://arxiv.org/pdf/0908.1803.pdf>.

Bản đồ họa thông tin đẹp mắt từ CERN tóm tắt các hạt cơ bản đã biết xem tại <http://tinyurl.com/cernparticle>.

Video độc đáo này về nguyên mẫu Orion phi hạt nhân minh họa ý tưởng về động cơ tên lửa với nhiên liệu là bom hạt nhân: <https://www.youtube.com/watch?v=E3Lxx2VAYi8>.

Đây là phần giới thiệu sơ phạm cho việc chèo thuyền laser: Robert L. Forward, “Roundtrip Interstellar Travel Using Laser-Pushed Lightsails,” *Journal of Spacecraft and Rockets* 21, no. 2

(tháng 3 – 4, 1984), truy cập trực tuyến tại <http://www.lunarsail.com/LightSail/rit-1.pdf>.

Jay Olson phân tích các nền văn hoá mở rộng vào vũ trụ trong “Homogeneous Cosmology with Aggressively Expanding Civilizations,” *Classical and Quantum Gravity* 32 (2015), truy cập trực tuyến tại <http://arxiv.org/abs/1411.4359>.

Phân tích khoa học đầy đủ đầu tiên về tương lai xa của chúng ta: Freeman J. Dyson, “Time Without End: Physics and Biology in an Open Universe,” *Reviews of Modern Physics* 51, no. 3 (1979): 447, truy cập trực tuyến tại http://blog.regehr.org/extra_files/dyson.pdf.

Công thức nói trên của Seth Lloyd cho chúng ta biết rằng thực hiện một phép tính toán trong một khoảng thời gian τ sẽ tốn mức năng lượng $E \geq h/4\tau$, trong đó h là hằng số Planck. Nếu chúng ta muốn thực hiện N phép tính liên tục (theo chuỗi) trong một khoảng thời gian T , vậy $\tau = TN$, vậy $E/N \geq hN/4T$, điều này cho thấy chúng ta có thể thực hiện các phép tính nối tiếp $N \leq 2\sqrt{ET}/h$ bằng cách sử dụng năng lượng E và thời gian T . Vì vậy cả năng lượng và thời gian là tài nguyên có càng nhiều càng có lợi. Nếu bạn chia năng lượng của bạn cho n phép tính toán song song khác nhau, chúng có thể chạy chậm hơn và hiệu quả hơn, với $N \leq 2\sqrt{ETn}/h$. Nick Bostrom ước tính rằng mô phỏng một cuộc sống 100 năm của con người đòi hỏi khoảng $N = 10^{27}$ phép tính. Nếu bạn muốn xem một bài lập luận cẩn thận về lý do nguồn gốc của sự sống có lẽ cần tới một tỉ lệ cực hiếm hơn, đặt những người hàng xóm gần nhất cách chúng ta hơn 101000 mét, tôi đề xuất video này của nhà vật lý học và nhà sinh vật học vũ trụ Princeton Edwin Turner: “Improbable Life: An Unappealing but Plausible Scenario for Life’s Origin on Earth,” tại <https://www.youtube.com/watch?v=Bt6n6Tu1beg>. Bài viết của Martin Rees về cuộc tìm kiếm trí tuệ ngoài trái đất: <https://www.edge.org/annual-question/2016/response/26665>.

Chương 7

Cuộc thảo luận phổ biến về công trình “thích nghi bởi tiêu tán” của Jeremy England có thể được tìm thấy tại Natalie Wolchover, “A New Physics Theory of Life,” *Scientific American* (28 tháng 1, 2014), truy cập trực tuyến tại <https://www.scientificamerican.com/article/a-new-physics-theory-of-life/>. Order Out of Chaos: Man’s New Dialogue with Nature (New York: Bantam, 1984) của Ilya Prigogine và Isabelle Stengers đặt nền móng cho nó.

Để biết thêm về cảm xúc và nguồn gốc sinh lý của chúng: William James, *Principles of Psychology* (New York: Henry Holt & Co., 1890); Robert Ornstein, *Evolution of Consciousness: The Origins of the Way We Think* (New York: Simon & Schuster, 1992); António Damásio, *Descartes’ Error: Emotion, Reason, and the Human Brain* (New York: Penguin, 2005); và António Damásio, *Self Comes to Mind: Constructing the Conscious Brain* (New York: Vintage, 2012).

Eliezer Yudkowsky bàn luận về việc định hướng các mục tiêu của AI thân thiện không phải theo mục tiêu hiện tại của chúng ta, mà theo ý chí ngoại suy nhất quán (coherent extrapolated volition - CEV). Nói một cách đơn giản nó được định nghĩa là những gì phiên bản lý tưởng của chúng ta sẽ muốn nếu chúng ta biết nhiều hơn, nghĩ nhanh hơn và giống với con người chúng ta muốn trở thành hơn. Yudkowsky bắt đầu chỉ trích CEV ngay sau khi xuất bản nó vào năm 2004

(<http://intelligence.org/files/CEV.pdf>), cả bởi lí do khó áp dụng cũng như không rõ liệu nó có thể đồng quy thành bất kỳ điều gì rõ ràng hay không.

Trong phương pháp học tăng cường nghịch đảo, ý tưởng cốt lõi là AI đang cố gắng tối đa tính thỏa mãn vì mục tiêu không phải của chính nó, mà là của chủ sở hữu con người. Do đó, nó được khuyến khích phải thận trọng khi không rõ chủ sở hữu muốn gì và cố gắng hết sức để tìm hiểu. Nó cũng nên thấy ổn khi chủ sở hữu tắt nó đi, vì điều đó có nghĩa là nó đã hiểu lầm những gì chủ sở hữu thực sự muốn.

Bài viết của Steve Omohundro về sự phát triển mục tiêu AI, “The Basic AI Drives,” có thể tìm thấy tại

<http://tinyurl.com/omohundro2008>. Xuất bản lần đầu trong Artificial General Intelligence 2008: Proceedings of the First AGI Conference, ed. Pei Wang, Ben Goertzel và Stan Franklin (Amsterdam: IOS, 2008), 483–492.

Cuốn sách gây nhiều suy ngẫm và tranh cãi về những gì xảy ra khi trí tuệ bị lợi dụng để mù quáng nghe theo mệnh lệnh mà không chất vấn tiêu chuẩn đạo đức: Hannah Arendt, Eichmann in Jerusalem: A Report on the Banality of Evil (New York: Penguin, 1963). Một nan đề tương tự xuất hiện trong đề xuất gần đây của Eric Drexler (<http://www.fhi.ox.ac.uk/reports/2015-3.pdf>), để giữ siêu trí tuệ trong tầm kiểm soát thì chia nó thành các phần đơn giản, mà không phần nào hiểu được toàn cảnh. Nếu cách này hiệu quả, nó sẽ cung cấp một công cụ cực kỳ mạnh mẽ không có định hướng đạo đức nội tại, thực hiện mọi ước muốn của chủ sở hữu mà không có bất kỳ lo ngại đạo đức nào. Điều này gợi nhớ đến bộ máy quan liêu bị chia nhỏ trong chế độ độc tài đầy khắc nghiệt: một bộ phận chế tạo vũ khí không biết tới việc chúng sẽ được sử dụng ra sao, bộ phận khác hành hình tù nhân mà không biết lí do họ bị kết án, vân vân.

Biến thể hiện đại của Quy tắc vàng là ý tưởng của John Rawls rằng một tình huống giả định là công bằng nếu không ai thay đổi nó mà không biết trước họ sẽ là người nào trong đó.

Ví dụ, chỉ số IQ của rất nhiều quan chức hàng đầu dưới quyền Hitler được biết khá cao. Xem “How Accurate Were the IQ Scores of the High-Ranking Third Reich Officials Tried at Nuremberg?,” Quora, truy cập tại <http://tinyurl.com/nurembergiq>.

Chương 8

Phần về ý thức của Stuart Sutherland khá thú vị: Macmillan Dictionary of Psychology (London: Macmillan, 1989).

Erwin Schrödinger, một trong số những người là cha đẻ của cơ học lượng tử, đã đưa ra nhận xét này trong cuốn sách Mind and Matter trong lúc suy ngẫm về quá khứ - và điều gì hẳn sẽ xảy ra nếu sự sống có ý thức chưa bao giờ phát triển ngay từ đầu. Mặt khác, sự lớn mạnh của AI làm tăng khả năng hợp lý chúng ta có

thể kết thúc ở vở kịch với những băng ghế trống không trong tương lai.

Bách khoa toàn thư Triết học Stanford đưa ra bản khảo sát rộng rãi về các định nghĩa và cách sử dụng khác nhau của từ “consciousness” (ý thức): <http://tinyurl.com/stanfordconsciousness>.

Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (New York: Harper- Collins, 2017): 116.

Đây là bản giới thiệu xuất sắc về Hệ thống 1 và Hệ thống 2 từ người tiên phong: Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus & Giroux, 2011).

Xem Christof Koch, *The Quest for Consciousness: A Neurobiological Approach* (New York: W. H. Freeman, 2004).

Có lẽ chúng ta chỉ biết một phần nhỏ (khoảng 10–50 bit) của thông tin đi vào bộ não chúng ta mỗi giây: K. Küpfmüller, 1962, “Nachrichtenverarbeitung im Menschen,” trong *Taschenbuch der Nachrichtenverarbeitung*, ed. K. Steinbuch (Berlin: Springer-Verlag, 1962): 1481–1502. T. Nørretranders, *The User Illusion: Cutting Consciousness Down to Size* (New York: Viking, 1991).

Michio Kaku, *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind* (New York: Doubleday, 2014); Jeff Hawkins và Sandra Blakeslee, *On Intelligence* (New York: Times Books, 2007); Stanislas Dehaene, Michel Kerszberg và Jean-Pierre Changeux, “A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks,” *Proceedings of the National Academy of Sciences* 95 (1998): 14529–14534.

Video kỷ niệm thí nghiệm nổi tiếng “Tôi có thể ngửi thấy mùi bánh mì nướng cháy” của Penfield: <https://www.youtube.com/watch?v=mSN86kphL68>.

Chi tiết về vỏ não giác quan vận động: Elaine Marieb và Katja Hoehn, *Anatomy & Physiology*, 3rd ed. (Upper Saddle River, NJ: Pearson, 2008), 391–395.

Nghiên cứu về mối tương quan thần kinh của ý thức (NCCs) đã trở nên khá phổ biến trong cộng đồng khoa học thần kinh những năm gần đây — tìm đọc, ví dụ, Geraint Rees, Gabriel Kreiman,

và Christof Koch, “Neural Correlates of Consciousness in Humans,” *Nature Reviews Neuroscience* 3 (2002): 261–270, và Thomas Metzinger, *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (Cambridge, MA: MIT Press, 2000).

Cách thức hoạt động của phương pháp riet tiêu chớp liên tục: Christof Koch, *The Quest for Consciousness: A Neurobiological Approach* (New York: W. H. Freeman, 2004); Christof Koch and Naotsugu Tsuchiya, “Continuous Flash Suppression Reduces Negative Afterimages,” *Nature Neuroscience* 8 (2005): 1096–1101.

Christof Koch, Marcello Massimini, Melanie Boly và Giulio Tononi, “Neural Correlates of Consciousness: Progress and Problems,” *Nature Reviews Neuroscience* 17 (2016): 307.

Xem Koch, *The Quest for Consciousness*, p. 260, và thảo luận sâu hơn tại *Stanford Encyclopedia of Philosophy*, <http://tinyurl.com/consciousnessdelay>.

Đồng bộ hóa nhận thức có ý thức: David Eagleman, *The Brain: The Story of You* (New York: Pantheon, 2015), và *Stanford Encyclopedia of Philosophy*, <http://tinyurl.com/consciousnesssync>.

Benjamin Libet, *Mind Time: The Temporal Factor in Consciousness* (Cambridge, MA: Harvard University Press, 2004); Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze và John-Dylan Haynes, “Unconscious Determinants of Free Decisions in the Human Brain,” *Nature Neuroscience* 11 (2008): 543–545, tại <http://www.nature.com/neuro/journal/v11/n5/full/nn.2112.html>.

Ví dụ về các phương pháp tiếp cận lý thuyết mới về ý thức:

Daniel Dennett, *Consciousness Explained* (Back Bay Books, 1992)

Bernard Baars, *In the Theater of Consciousness: The Workspace of the Mind* (New York: Oxford University Press, 2001)

Christof Koch, *The Quest for Consciousness: A Neurobiological Approach* (New York: W. H. Freeman, 2004)

Gerald Edelman and Giulio Tononi, *A Universe of Consciousness: How Matter Becomes Imagination* (New York: Hachette, 2008)

António Damásio, *Self Comes to Mind: Constructing the Conscious Brain* (New York: Vintage, 2012)

Stanislas Dehaene, *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (New York: Viking, 2014)

Stanislas Dehaene, Michel Kerszberg and Jean-Pierre Changeux, "A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks," *Proceedings of the National Academy of Sciences* 95 (1998): 14529–14534

Stanislas Dehaene, Lucie Charles, Jean-Rémi King and Sébastien Marti, "Toward a Computational Theory of Conscious Processing," *Current Opinion in Neurobiology* 25 (2014): 760–784

Thảo luận sâu về các cách sử dụng khác nhau của thuật ngữ "emergence" (đột sinh) trong vật lý và triết học bởi David Chalmers: <http://cse3521.artifice.cc/Chalmers-Emergence.pdf>. Tôi đang tranh luận rằng ý thức là cách thông tin cảm nhận khi được xử lý theo những cách phức tạp nhất định:

<https://arxiv.org/abs/physics/0510188>, <https://arxiv.org/abs/0704.0646>, Max Tegmark, *Our Mathematical Universe* (New York: Knopf, 2014). David Chalmers bộc lộ quan điểm tương tự trong cuốn sách năm 1996 của mình *The Conscious Mind* (Trí óc có ý thức): "Kinh nghiệm là thông tin từ bên trong; vật lý là thông tin từ bên ngoài."

Adenauer Casali et al., "A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior," *Science Translational Medicine* 5 (2013): 198ra105, tại <http://tinyurl.com/zapzip>.

Lý thuyết thông tin tích hợp không phù hợp với hệ thống liên tục:

<https://arxiv.org/abs/1401.1219>

<http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00063/full>

<https://arxiv.org/abs/1601.02626>

Phỏng vấn với Clive Wearing, người trí nhớ ngắn hạn chỉ có 30 giây: <https://www.youtube.com/watch?v=WmzU47i2xgw>.

Phê bình về IIT của Scott Aaronson:

<http://www.scottaaronson.com/blog/?p=1799>.

Phê bình về IIT của Cerrullo, lập luận rằng tích hợp không phải điều kiện đủ cho ý thức: <http://tinyurl.com/cerrullocritique>.

Dự đoán IIT rằng con người mô phỏng sẽ là xác sống:

<http://rstb.royalsocietypublishing.org/content/370/1668/20140167>.

Phê bình về IIT của Shanahan:

<https://arxiv.org/pdf/1504.05696.pdf>.

Thị lực mù: <http://tinyurl.com/blindsight-paper>.

Có lẽ chúng ta chỉ biết một phần nhỏ (khoảng 10–50 bit) của thông tin đi vào bộ não chúng ta mỗi giây: Küpfmüller, “Nachrichtenverarbeitung im Menschen”; Nørretranders, *The User Illusion*.

Trường hợp ủng hộ và phản đối “ý thức không truy cập”: Victor Lamme, “How Neuroscience Will Change Our View on Consciousness,” *Cognitive Neuroscience* (2010): 204–220, truy cập tại

<http://www.tandfonline.com/doi/abs/10.1080/17588921003731586>.

“Selective Attention Test,” tại <https://www.youtube.com/watch?v=vJG698U2Mvo>.

Xem Lamme, “How Neuroscience Will Change Our View on Consciousness,” n. 28.

Vấn đề này và các thứ liên quan khác được thảo luận chi tiết trong cuốn *Consciousness Explained* của Daniel Dennett.

Xem Kahneman, *Thinking, Fast and Slow*, trích dẫn n. 5.

Bách khoa toàn thư Triết học Stanford nhìn nhận các tranh luận xung quanh ý chí tự do: <https://plato.stanford.edu/entries/freewill>.

Video về Seth Lloyd giải thích tại sao AI cảm thấy như thể nó có ý chí tự do: <https://www.youtube.com/watch?v=Epj3DF8jDWk>.

Xem Steven Weinberg, *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature* (New York: Pantheon, 1992).

Phân tích khoa học đầy đủ đầu tiên về tương lai xa của chúng ta: Freeman J. Dyson, “Time Without End: Physics and Biology in an Open Universe,” *Reviews of Modern Physics* 51, no. 3

(1979): 447, truy cập trực tuyến tại http://blog.regehr.org/extra_files/dyson.pdf.

Phần kết

Bức thư ngỏ (<http://futureoflife.org/ai-open-letter>) đến từ hội nghị Puerto Rico lập luận rằng nghiên cứu về cách thức khiến hệ thống AI ổn định và hữu ích vừa quan trọng vừa hợp thời, và đã có những hướng nghiên cứu vững chắc có thể được theo đuổi từ bây giờ, như được minh họa trong tài liệu về ưu tiên nghiên cứu này:

http://futureoflife.org/data/documents/research_priorities.pdf.

Video tôi phỏng vấn Elon Musk về an toàn AI có thể tìm thấy trên YouTube tại <https://www.youtube.com/watch?v=rBw0eoZTY-g>.

Đây là một video tổng hợp hay ho về gần như mọi nỗ lực hạ cánh của tên lửa SpaceX, đỉnh điểm là lần đầu tiên hạ cánh thành công trên đại dương: <https://www.youtube.com/watch?v=AllaFzIPaG4>.

Elon Musk đăng trên twitter về cuộc thi nhằm tài trợ an toàn AI của chúng tôi: <https://twitter.com/elonmusk/status/555743387056226304>.

Elon Musk đăng trên twitter về bức thư ngỏ ủng hộ an toàn AI của chúng tôi: <https://twitter.com/elonmusk/status/554320532133650432>.

Erik Sofge trong “An Open Letter to Everyone Tricked into Fearing Artificial Intelligence” (Popular Science, 14 tháng 1, 2015) chọc cười về tin tức gây lo lắng trong bức thư ngỏ của chúng tôi: <http://www.popsci.com/open-letter-everyone-tricked-fearing-ai>.

Elon Musk đăng trên twitter về khoản đóng góp lớn của mình cho Viện Tương lai Sự sống và cho thế giới của các nhà nghiên cứu an toàn AI: <https://twitter.com/elonmusk/status/555743387056226304>.

Để biết thêm về Quan hệ đối tác AI nhằm mang lại lợi ích cho mọi người và xã hội, hãy xem trang web của họ: <https://www.partnershiponai.org>.

Một vài ví dụ về các báo cáo gần đây với ý kiến về AI: One Hundred Year Study on Artificial Intelligence, Report of the 2015 Study Panel, “Artificial Intelligence and Life in 2030” (tháng 9, 2016), tại <http://tinyurl.com/stanfordai>; báo cáo của Nhà Trắng về tương lai của AI: <http://tinyurl.com/obamaAIreport>; báo cáo của Nhà Trắng về AI và việc làm: <http://tinyurl.com/AIjobsreport>; Báo cáo của IEEE về AI và hạnh phúc con người, “Ethically Aligned Design, Version 1” (13 tháng 12, 2016), tại http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf; bản đồ cho ngành rô-bốt Hoa Kỳ: <http://tinyurl.com/roboticsmap>. Trong các nguyên tắc không được chọn, một trong số yêu thích của tôi là: “Ý thức thận trọng: Nếu không có sự nhất trí, chúng ta nên tránh những giả định chắc chắn về việc liệu AI tiên tiến sở hữu hay yêu cầu có ý thức hay cảm xúc.” Nó đã trải qua nhiều vòng, và cuối cùng, từ “ý thức” gây tranh cãi đã được thay thế bằng từ “kinh nghiệm chủ quan” - nhưng dù vậy nguyên tắc này chỉ được 88% chấp thuận, hơi thấp hơn một chút so với mức yêu cầu 90%.

Thảo luận về siêu trí tuệ cùng với Elon Musk và những bộ não tuyệt vời khác: <http://tinyurl.com/asilomarAI>.

11. SSD: Ổ cứng điện tử thể rắn

12. DIMM: Mô-đun bộ nhớ hàng đôi

13. SIMM: Mô-đun bộ nhớ hàng một

14. Người Amish là một nhóm tín hữu Kitô giáo duy truyền thống. Đến tận bây giờ, họ vẫn duy trì cuộc sống đơn giản với các quy tắc nghiêm ngặt, theo cách tự cung tự cấp và tránh xa công nghệ hiện đại.

[←1]

AI: Artificial Intelligence

[←2]

Ý chỉ việc dùng chính AI này để phát triển ra AI tiếp theo vượt bậc hơn.

[←3]

Để cho đơn giản, tôi dùng tình hình kinh tế và công nghệ hiện tại cho câu chuyện này, tuy rằng phần lớn các nhà nghiên cứu dự đoán rằng AI tổng quát có trình độ như con người còn cách ta ít nhất vài thế kỷ nữa. Kế hoạch Omega sẽ được thực hiện dễ hơn nhiều trong tương lai nếu nền kinh tế kỹ thuật tiếp tục phát triển và càng nhiều dịch vụ có thể đặt qua mạng mà không ai còn thắc mắc về nó.

[←4]

Dịch vụ này được đặt theo tên cỗ máy chơi cờ nổi tiếng có ngoại hình một người Thổ Nhĩ Kỳ, chế tạo năm 1770.

[←5]

Human Intelligence Tasks

[←6]

Chứng khoán phái sinh (Derivatives): là các công cụ tài chính có giá trị phụ thuộc vào giá trị của các tài sản cơ sở (tài sản, chỉ số, lãi suất).

[←7]

Có lẽ nói đến cuộc khủng hoảng thị trường chứng khoán năm 2015-16. Như tác giả chú thích ở trên, câu chuyện giả tưởng này chưa xảy ra, nên các sự kiện chỉ là ví dụ minh họa.

[←8]

Podcast: là kênh chương trình dài kỳ với nội dung có thể ở dạng âm thanh, video, hay tệp PDF mà người nghe có thể tải về máy. Đây là hệ thống phân phối nội dung do hãng Apple phát triển. Từ podcast được tạo nên bằng cách kết hợp hai từ là iPod (một thiết bị của Apple) và “broadcast” (phát sóng).

[←9]

Từ gốc: Hair-triggered alert – một chính sách quân sự của Mỹ có từ thời Chiến tranh lạnh, cho phép phóng nhanh vũ khí hạt nhân. Tên lửa được lắp trên bộ phận kích hoạt rất nhạy, luôn được

duy trì ở trạng thái sẵn sàng phát động và có thể được phóng trong thời gian chưa đến 10 phút.

[←10]

Tại sao sự sống trở nên phức tạp hơn? Thường quá trình tiến hóa sẽ tưởng thưởng cho những dạng sống đủ phức tạp để có thể dự đoán và khai thác các quy luật trong môi trường của nó, vì thế trong một môi trường phức tạp hơn, sự sống phức tạp và thông minh hơn sẽ phát triển. Sự sống thông minh hơn này cũng tạo ra một môi trường phức tạp hơn cho các dạng sống cạnh tranh, tới lượt các dạng này cũng tiến hóa phức tạp hơn, cuối cùng tạo ra một hệ sinh thái với sự sống cực kỳ phức tạp.

[←11]

Đơn vị thông tin lưu trữ máy tính. 1 terabyte = 1.024 gigabyte

[←12]

Artificial General Intelligence

[←13]

Massachusetts Institute of Technology (Viện Công nghệ Massachusetts): Viện đại học nghiên cứu tư thực nổi tiếng về công nghệ ở Mỹ

[←14]

Singularity – Điểm kỳ dị (công nghệ): giả thuyết về thời điểm trí tuệ nhân tạo cùng các công nghệ khác phát triển tới mức xã hội loài người phải trải qua một sự thay đổi khó lường. Một khi thời điểm kỳ dị xuất hiện, các sự kiện sau đó sẽ không thể dự đoán và tiên lượng được nữa.

[←15]

Cuộc đối thoại AI rất quan trọng cả về tính cấp thiết lẫn tầm ảnh hưởng. So với biến đổi khí hậu, thứ có thể tàn phá 50 đến 200 năm tới, nhiều chuyên gia mong đợi AI có tác động còn lớn hơn trong vài thập kỷ tới và có khả năng mang đến cho chúng ta công nghệ để giảm thiểu biến đổi khí hậu. So với chiến tranh,

khủng bố, thất nghiệp, nghèo đói, di cư và các vấn đề về công bằng xã hội, sự trỗi dậy của AI sẽ có tác động tổng thể lớn hơn – thực vậy, chúng ta sẽ tìm hiểu trong cuốn sách này về cách nó có thể thống trị những gì xảy ra với tất cả các vấn đề đó, dù tốt hay xấu.

[←16]

Để thấy rõ điều này, hãy tưởng tượng bạn sẽ phản ứng ra sao khi có người khẳng định khả năng đạt được các thành tựu thể thao ở trình độ Thế vận hội Olympic có thể đo đếm được bằng một con số gọi là “chỉ số vận động” (Athletic Quotient – AQ), và như thể vận động viên Olympic có AQ cao nhất sẽ giành huy chương vàng tất cả các môn thể thao.

[←17]

Một số người thích dùng “AI ở trình độ con người” hay “AI mạnh” để thay thế cho AGI, nhưng cả hai từ này đều có vấn đề. Ngay cả một máy tính bỏ túi cũng là AI ở trình độ con người theo nghĩa hẹp. Còn từ trái nghĩa của “AI mạnh” là “AI yếu”, nhưng sẽ thật kỳ quặc khi gọi những hệ thống AI hẹp như Deep Blue, Watson, và AlphaGo là “yếu”.

[←18]

RNA: một axit nucleic tương tự DNA

[←19]

NAND là viết tắt của NOT AND: Một cổng AND cho đầu ra là 1 chỉ khi đầu vào thứ nhất là 1 và đầu ra thứ hai là 1, nên NAND cho đầu ra trái ngược với AND.

[←20]

0

[←21]

Dành cho bạn nếu bạn thích môn toán, hai lựa chọn phổ biến của hàm số kích hoạt gọi là hàm chữ S $\sigma(x) \equiv 1/(1 + e^{-x})$ và hàm con dốc $\sigma(x) = \max\{0, x\}$, mặc dù đã chứng minh rằng hầu như

bất kỳ hàm nào cũng được miễn là nó không phải hàm tuyến tính (đồ thị là đường thẳng). Mô hình nổi tiếng của Hopfield sử dụng $\sigma(x) = -1$ nếu $x < 0$ và $\sigma(x) = 1$ nếu $x \geq 0$. Nếu các trạng thái nơ-ron được lưu trữ trong một vec-tơ, thì mạng nơ-ron được cập nhật đơn giản bằng cách nhân vec-tơ với ma trận chứa các cặp đôi xi-nap và ứng dụng hàm σ cho tất cả các nhân tố.

[←22]

Bảng xếp hạng tôn vinh những người trẻ dưới 30 tuổi có thành tựu xuất sắc trong nhiều lĩnh vực do tạp chí Forbes bình chọn

[←23]

DeepMind là một công ty trí tuệ nhân tạo của Anh được thành lập vào năm 2010, sau đó được Google mua lại vào năm 2014.

[←24]

Atari là một hãng lập trình trò chơi điện tử và máy tính gia đình tại Mỹ, thành lập từ năm 1972. Sau vụ sụp đổ thị trường trò chơi ở Bắc Mỹ vào 1983, công ty này đã đóng cửa và bị chia tách, bán lại nhiều lần.

[←25]

Nếu bạn muốn một bản đồ chi tiết hơn về tình hình nghiên cứu an toàn AI, có một bản tương tác ở đây, được cộng đồng nỗ lực phát triển dẫn đầu bởi Richard Mallah của FLI: <https://futureoflife.org/landscape>.

[←26]

Nguyên văn “the blue screen of death” và “the spinning wheel of doom”: tình trạng xuất hiện trên màn hình của 2 hệ điều hành Windows và Mac OS khi gặp sự cố.

[←27]

High-assurance cyber military systems

[←28]

Nói cho chuẩn hơn thì kiểm định sẽ đặt câu hỏi liệu một hệ thống có đáp ứng các đặc tả của nó hay không, trong khi thẩm định hỏi xem liệu các đặc tả chính xác có được lựa chọn không.

[←29]

Kể cả khi bao gồm cả sự cố này trong số liệu thống kê, Autopilot của Tesla đã giảm 40% sự cố khi hoạt động:
<http://tinyurl.com/teslasafety>.

[←30]

Mạng Toàn Cầu

[←31]

Chụp cộng hưởng từ chức năng