

NICK BOSTROM

SIÊU TRÍ TUỆ

SUPERINTELLIGENCE

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

PHẠM HỒNG ANH - NGUYỄN DUY ANH dịch



NEW YORK TIMES BESTSELLER

"Chúng ta nên cẩn
thận với trí tuệ nhân
tạo... Nó còn nguy
hiểm hơn cả vũ khí
hạt nhân."

ELON MUSK

alpha
books

ets

Education
Technology
Science



Nhà xuất bản
thế giới

THÔNG TIN EBOOK



Siêu Trí Tuệ

Nick Bostrom

Nhiều dịch giả Dịch

Nhà Xuất Bản Thế Giới

More Books At

fb.com/groups/yeukindlevietnam

Notion Ebook Member

tinyurl.com/downloadebookyeukindle



Sách này không làm chú thích ^^
tinyurl.com/kindlemagzone

Mục lục

Lời giới thiệu

Câu chuyện ngụ ngôn dorr dang về đàn chim sẻ

Lời nói đầu

Lời cảm ơn

Chương 1: Sự phát triển trong quá khứ và những năng lực hiện có

Chương 2: Con đường đến với siêu trí tuệ

Chương 3: Các dạng thức siêu trí tuệ

Chương 4: Động lực của sự bùng nổ trí tuệ

Chương 5: Lợi thế cạnh tranh quyết định

Chương 6: Siêu năng lực nhận thức

Chương 7: Ý chí siêu trí tuệ

Chương 8: Kết quả mặc định có phải là ngày tận thế?

Chương 9: Vấn đề kiểm soát

Chương 10: Tiên tri, thần đèn, toàn năng, công cụ

Chương 11: Những kịch bản đa cực

Chương 12: Thụ đắc các giá trị

Chương 13: Lựa chọn tiêu chí chọn lựa

Chương 14: Bức tranh chiến lược

Chương 15: Giai đoạn nước rút

Lời kết

Chú thích

Tài liệu tham khảo

Giải thích một số thuật ngữ

Chỉ mục

Lời giới thiệu

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

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

loại công nghệ với robot thông minh, công nghệ đa tác tử tự trị và hợp tác, lập luận thời gian thực với dữ liệu từ nhiều cảm biến... Dù đã đạt được những kết quả khích lệ, vẫn còn một chặng đường rất dài để có thể đạt được khát vọng về AI ở cấp độ con người.

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

Khi nhận xét rằng các phương pháp của điều khiển học tập trung vào xử lý tín hiệu và ra quyết định cho những công việc có thể phức tạp nhưng không đòi hỏi quá nhiều năng lực trí tuệ cao của con người. Ví dụ như một chiếc xe tự lái chủ yếu cần nhận ra mọi thứ trên con đường, biết đi nhanh đi chậm, biết tránh các xe khác để đi an toàn. Nhiều thành tựu nổi bật khoảng hai mươi năm qua dưới tên “AI” đạt được chủ yếu trong các lĩnh vực như nhận dạng và điều khiển chuyển động, hay thống kê và học máy thống kê nhằm tìm kiếm các mẫu dạng trong dữ liệu và đưa ra các dự đoán có căn cứ,

kiểm tra các giả thuyết và đề xuất các quyết định. Các hệ thống tìm kiếm tài liệu, phân loại văn bản, phát hiện gian lận, khuyến nghị hành động, cá thể hoá việc chẩn đoán và chữa bệnh, phân tích mạng xã hội... là những thành công lớn của AI và được theo đuổi bởi các công ty khổng lồ như Google, Netflix, Facebook và Amazon... Những thành công này cho ta chứng kiến trong hai thập kỷ qua nhiều tiến bộ về học thuật và phát triển sản phẩm cho một hướng phát triển nữa của AI với tên gọi “AI tăng cường trí tuệ”. Hướng nghiên cứu này của AI giúp con người giải quyết những bài toán, những nhiệm vụ cụ thể cần đến trí thông minh và sự sáng tạo. Những nhiệm vụ này có thể có ở mọi lĩnh vực hoạt động của con người, từ nông nghiệp, du lịch, thương mại đến y học, giáo dục... Hướng phát triển này được gọi là “AI chuyên dụng” (narrow AI) hoặc “AI yếu” (weak AI). “Yếu” ở đây vì không làm việc vạn năng, nhưng rất “mạnh” ở việc cụ thể phải làm.

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

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

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

Câu chuyện ngụ ngôn dở dang về đàn chim sẻ

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

“Chúng ta thật nhỏ bé và yếu đuối. Nghĩ mà xem, cuộc sống sẽ dễ dàng làm sao nếu chúng ta có một cậu cú mèo để giúp xây tổ nhỉ!” “Phải rồi!” Một con khác nói. “Và mình có thể nhờ cậu ấy chăm sóc các cụ sẻ già và cả lũ sẻ non nữa chứ.”

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

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

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

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

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

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

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

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

Lời nói đầu

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

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

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

đã tồn tại, nó sẽ ngăn cản chúng ta thay thế nó hay thay đổi những ưu tiên của nó, và khi đó, số phận nhân loại sẽ bị định đoạt.

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

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

Đây không phải là một cuốn sách dễ viết. Tôi hy vọng rằng con đường đã vạch ra sẽ cho phép những nhà thám hiểm khác vươn tới những ranh giới mới nhanh chóng và dễ dàng hơn, để khi đến đó, họ

vẫn còn khỏe khoắn và sẵn sàng tham gia vào việc tiếp tục mở rộng tầm hiểu biết của chúng ta. (Và nếu con đường đã mở có chút gập ghềnh quanh co, tôi hy vọng rằng những người quan sát, khi đánh giá kết quả, sẽ không coi nhẹ những gian truân sẽ gặp phải trên con đường đó!)

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

Nhiều điểm đưa ra trong cuốn sách này có thể là không đúng.²

Cũng có thể tôi chưa cân nhắc đến một số điểm trọng yếu, và do đó một số hay toàn bộ kết luận đã trở nên vô giá trị. Tôi đã khá cố gắng để chỉ ra nhiều mức độ không chắc chắn trong suốt cuốn sách và rào đón chúng bằng những từ như “dường như”, “có lẽ”, “có khả năng”, “gần như”, “có thể”, “rất có thể”, “gần như chắc chắn”. Các từ hạn định được sử dụng vô cùng thận trọng và dè dặt, mặc dù vậy, sự khiêm tốn về nhận thức này vẫn chưa đủ mà còn cần được bổ trợ

bằng việc chấp nhận một cách có hệ thống sự bất định và khả năng xảy ra sai lầm. Đây không phải là sự khiêm tốn giả tạo: trong khi tin rằng cuốn sách của mình có thể đã sai và lạc lối nghiêm trọng, tôi đã nghĩ rằng những quan điểm khác được trình bày trong đó còn tệ hại hơn – bao gồm cả quan điểm mặc định, hay “giả thuyết không” (null hypothesis), mà theo đó, chúng ta tạm thời có thể lờ đi viễn cảnh về siêu trí tuệ một cách an toàn và hợp lý.

Lời cảm ơn

Lớp “màng” bao quanh quá trình viết lách là khá mỏng manh và dễ “thấm”. Nhiều khái niệm và ý tưởng được tạo ra trong khi viết cuốn sách này đã được “rò rỉ” ra ngoài và trở thành một phần của các cuộc tranh luận rộng hơn; và tất nhiên, nhiều cái nhìn sâu sắc nảy sinh từ bên ngoài đã được đưa vào trong quá trình viết sách. Tôi đã cố gắng sử dụng các công cụ trích dẫn, nhưng những ý kiến có tầm ảnh hưởng là quá nhiều nên khó mà ghi lại được đầy đủ.

Vì những cuộc trao đổi kỹ càng giúp đả thông tư tưởng, tôi xin được gửi lời cảm ơn đến nhiều người, bao gồm: Sam Altman, Dario Amodei, Ross Andersen, Stuart Armstrong, Owen Cotton-Barratt, Nick Beckstead, Yoshua Bengio, David Chalmers, Paul Christiano, Milan Ćirković, Andrew Critch, Daniel Dennett, David Deutsch, Daniel Dewey, Thomas Dietterich, Eric Drexler, David Duvenaud, Peter Eckersley, Amnon Eden, Oren Etzioni, Owain Evans, Benja Fallenstein, Alex Flint, Carl Frey, Zoubin Ghahramani, Ian Goldin, Katja Grace, Roger Grosse, Tom Gunter, J. Storrs Hal, Robin Hanson, Demis Hassabis, Geoffrey Hinton, James Hughes, Marcus Hutter, Garry Kasparov, Marcin Kulczycki, Patrick LaVictoire, Shane Legg, Moshe Looks, William MacAskill, Eric Mandelbaum, Gary Marcus, James Martin, Lillian Martin, Roko Mijic, Vincent Mueller, Elon Musk, Seán Ó hÉigeartaigh, Christopher Olah, Toby Ord, Laurent Orseau, Michael Osborne, Larry Page, Dennis Pamlin, Derek

Parfit, David Pearce, Huw Price, Guy Ravine, Martin Rees, Bill Roscoe, Francesca Rossi, Stuart Russell, Anna Salamon, Lou Salkind, Anders Sandberg, Julian Savulescu, Jürgen Schmidhuber, Bart Selman, Nicholas Shackel, Murray Shanahan, Noel Sharkey, Carl Shulman, Peter Singer, Nate Soares, Dan Stoicescu, Mustafa Suleyman, Jaan Tallinn, Alexander Tamas, Jessica Taylor, Max Tegmark, Roman Yampolskiy, và Eliezer Yudkowsky.

Vì những ý kiến đặc biệt chi tiết, tôi cảm ơn Milan Ćirković, Daniel Dewey, Owain Evans, Nick Hay, Keith Mansfield, Luke Muehlhauser, Toby Ord, Jess Riedel, Anders Sandberg, Murray Shanahan, và Carl Shulman. Và tôi cũng muốn cảm ơn Stuart Armstrong, Daniel Dewey, Eric Drexler, Alexandre Erler, Rebecca Roache, và Anders Sandberg vì sự tư vấn hoặc trợ giúp nghiên cứu trong những phần khác nhau.

Tôi cảm ơn Caleb Bell, Malo Bourgon, Robin Brandt, Lance Bush, Cathy Douglass, Alexandre Erler, John King, Kristian Rönn, Susan Rogers, Kyle Scott, Andrew SnyderBeattie, Cecilia Tilli, và Alex Vermeer vì đã giúp chuẩn bị bản thảo. Tôi muốn đặc biệt cảm ơn biên tập viên của mình là Keith Mansfield vì sự động viên của anh trong suốt quá trình thực hiện dự án.

Tôi xin lỗi tất cả những ai xứng đáng nhưng chưa được nhắc tên. Cuối cùng, tôi xin gửi lời cảm ơn chân thành đến các nhà tài trợ, bạn bè và gia đình: không có sự trợ giúp của mọi người công trình này đã không thể hoàn thành.

Chương 1

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

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

Các phương thức tăng trưởng và lịch sử nói chung

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

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

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

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

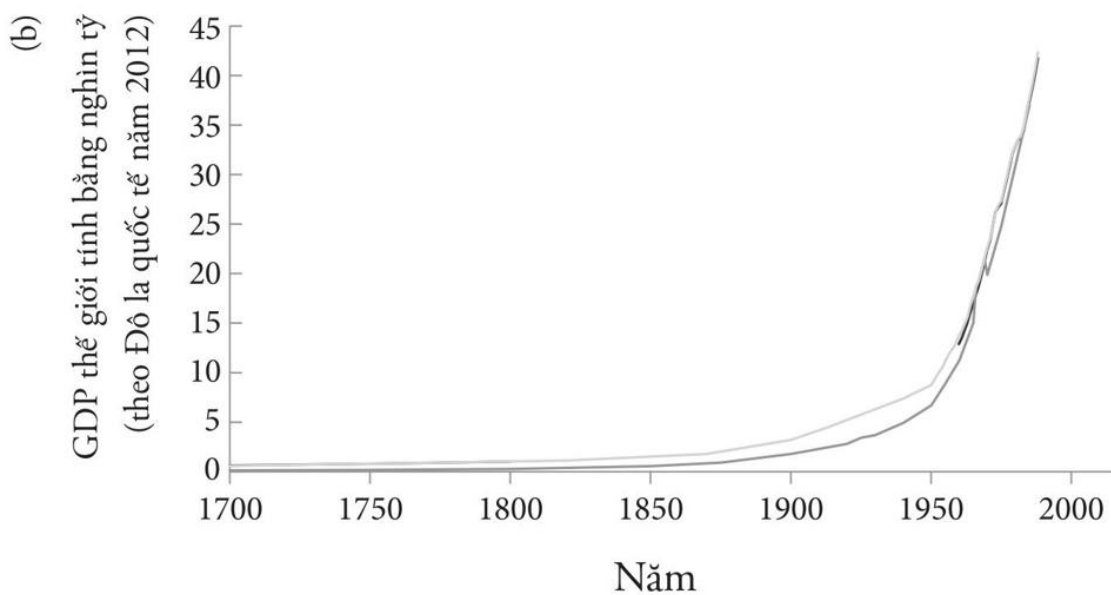
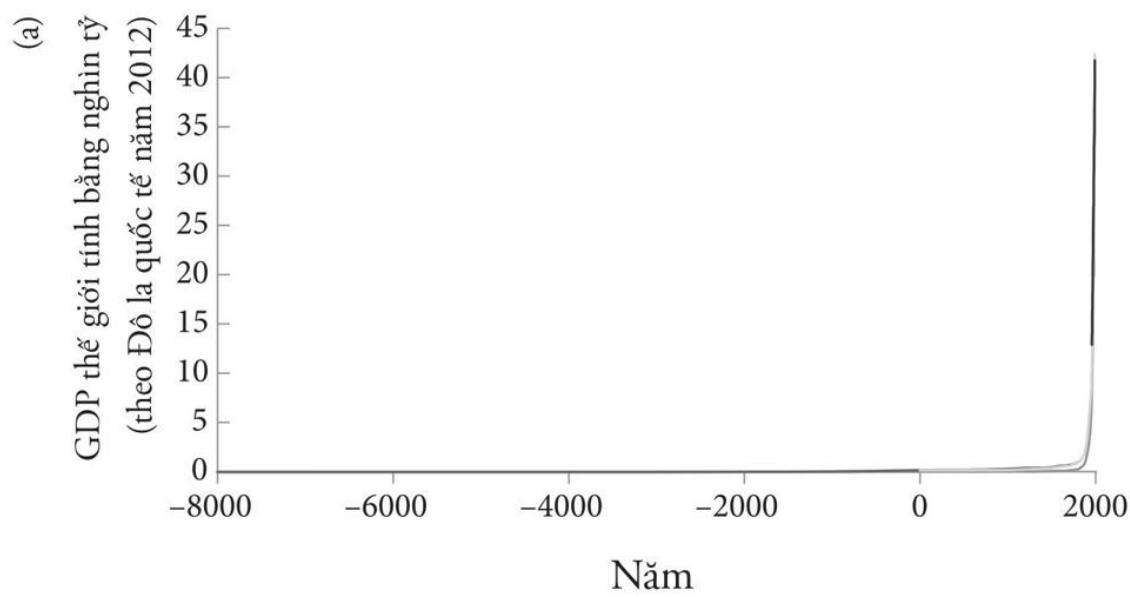
Thậm chí tốc độ tăng trưởng hiện tại cũng sẽ cho những kết quả ấn tượng nếu được duy trì trong một thời gian dài. Nếu nền kinh tế tiếp tục phát triển với tốc độ giống như 50 năm vừa qua thì thế giới sẽ giàu hơn hiện tại 4,8 lần vào năm 2050 và 34 lần vào năm 2100.²

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

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

Ý tưởng về một điểm kỳ dị công nghệ sắp xuất hiện đang được phổ biến rất rộng rãi, khởi đầu bằng tiểu luận quan trọng của Vernor

Vinge và tiếp tục với các tác phẩm của Ray Kurzweil và những người khác.⁴ Tuy nhiên thuật ngữ “điểm kỳ dị” đang được sử dụng lẫn lộn với nhiều ý nghĩa khác nhau, tạo ra một mớ nghĩa bóng huyễn hoặc nhưng vẫn ra vẻ hàn lâm và mang đầy tính ảo tưởng công nghệ.⁵ Do hầu hết các nghĩa này không liên quan đến luận điểm của chúng tôi nên từ “điểm kỳ dị” đã được loại bỏ để sử dụng thuật ngữ khác chính xác hơn.



Hình 1: Lịch sử dài hạn của GDP thế giới. Được vẽ bằng thang đo tuyến tính, lịch sử của kinh tế thế giới giống như một đường thẳng bám lấy trục x cho tới khi bất thành linh tăng vọt. (a) Kể cả khi chúng ta nhìn vào 10.000 năm gần đây nhất thì đồ thị gần như cũng đi lên theo chiều thẳng đứng. (b) Chỉ trong khoảng 100 năm qua, đường

cong đồ thị vượt hẳn lên trên mức không. (Những đồ thị khác nhau trên hình tương ứng với những bộ dữ liệu khác nhau, cho ra những đánh giá khác nhau đôi chút.⁶⁾

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

Những kỳ vọng lớn lao

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

tuệ nhân tạo tổng quát (artificial general intelligence) vẫn tin rằng máy móc thông minh còn cách chúng ta khoảng hai thập kỷ nữa.⁸

Hai thập kỷ là khoảng thời gian ưa thích của những “nhà tiên tri” về các thay đổi toàn diện: đủ gần để thu hút sự chú ý và có tính liên quan, nhưng lại đủ xa để giả định rằng một chuỗi đột phá mà hiện tại chỉ có thể hình dung một cách mờ nhạt có thể sẽ xảy ra. Hãy đối chiếu nó với các thang thời gian ngắn hơn: hầu hết các công nghệ sẽ có tác động lớn trên thế giới trong năm hay 10 năm tới hiện đều đang được sử dụng một cách hạn chế, còn những công nghệ sẽ định hình lại thế giới trong thời gian dưới 15 năm có thể cũng đã tồn tại dưới dạng các nguyên mẫu trong phòng thí nghiệm. 20 năm có thể cũng gần với khoảng thời gian còn lại diễn hình trong sự nghiệp của các nhà dự báo, gắn liền với rủi ro bị mất uy tín khi đưa ra một dự đoán táo bạo.

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

Trong chương tiếp theo, chúng tôi sẽ xem xét những con đường khác nhau có thể dẫn đến một thứ trí tuệ máy ở cấp độ con người. Nhưng xin lưu ý ngay từ đầu rằng cho dù có bao nhiêu điểm dừng từ

nay đến khi có được thứ trí tuệ máy đó thì cũng vẫn chưa phải là đích đến cuối cùng. Điểm dừng tiếp theo, chỉ một quãng đường ngắn nữa, là trí tuệ máy cấp độ siêu nhân. Chuyến tàu này có thể không dừng lại, thậm chí là không giảm tốc khi đến Ga Người (Humanville) mà sẽ lao nhanh qua nó.

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

Hãy định nghĩa một cỗ máy siêu thông minh (ultraintelligent) là cỗ máy có thể vượt qua mọi hoạt động trí tuệ của bất cứ người nào dù thông minh đến đâu. Vì việc thiết kế máy móc là một trong những hoạt động trí tuệ đó, nên một cỗ máy siêu thông minh thậm chí có thể thiết kế những chiếc máy tốt hơn; và không còn nghi ngờ gì nữa, khi đó sẽ có một “sự bùng nổ trí tuệ”. Khi đó, trí tuệ con người sẽ bị tụt hậu rất xa. Vì vậy, cỗ máy siêu thông minh đầu tiên là phát minh cuối cùng mà con người cần tạo ra, với điều kiện cỗ máy đủ dễ bảo để nói với chúng ta cách thức kiểm soát chính nó.¹⁰

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

máy có trí tuệ cấp độ con người, đến mức không thể nắm bắt được hậu quả của điều này – rằng máy móc sẽ dần trở thành siêu trí tuệ.

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

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

Mùa của hy vọng và tuyệt vọng

Mùa hè năm 1956 tại Đại học Dartmouth, 10 nhà khoa học cùng quan tâm đến mạng lưới thần kinh, lý thuyết automata và nghiên cứu trí tuệ cùng tổ chức một hội thảo sáu tuần. Dự án Mùa hè Dartmouth (Dartmouth Summer Project) này thường được coi như buổi bình minh của trí tuệ nhân tạo với tư cách là một ngành nghiên cứu. Nhiều người tham gia dự án sau này được công nhận là những nhân vật sáng lập. Viễn cảnh lạc quan của các đại biểu được thể hiện trong đề xuất trình lên Quỹ Rockefeller, đơn vị tài trợ cho sự kiện:

Chúng tôi đề nghị thực hiện một nghiên cứu về trí tuệ nhân tạo kéo dài trong hai tháng với 10 người... Nghiên cứu này được thực

hiện trên cơ sở phỏng đoán rằng mọi khía cạnh của tính năng học tập, hay bất cứ tính năng nào khác của trí tuệ, về nguyên tắc có thể được mô tả chính xác đến mức tạo ra được một cỗ máy để mô phỏng nó. Chúng tôi sẽ cố gắng tìm hiểu cách thức máy móc sử dụng ngôn ngữ, tạo dựng các khái niệm trừu tượng và cụ thể, giải quyết các vấn đề hiện tại chỉ thuộc về con người và tự cải thiện bản thân. Chúng tôi cho rằng có thể có tiến bộ đáng kể ở một hoặc nhiều vấn đề trong số đó nếu một nhóm các nhà khoa học được lựa chọn cẩn thận cùng làm việc trong một mùa hè.

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

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

chứng minh mãi mãi hơn nhiều so với nguyên bản, nhờ đó phê phán luận điểm cho rằng máy móc chỉ có thể “tư duy bằng con số” và cho thấy máy móc còn có khả năng suy luận và tạo ra các phép chứng minh logic.¹³ Một chương trình theo sau đó là General Problem Solver (Trình giải toán tổng quát) về nguyên tắc có thể giải một số lượng lớn các bài toán được xác định bằng công thức.¹⁴

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

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

^b *Nhãn-tưởng, chơi chữ với idea – ý tưởng. (BTV)*

Những phương pháp thành công trong các hệ thống trình diễn sơ khai này thường rất khó mở rộng cho nhiều dạng bài toán hoặc cho các bài toán khó hơn. Một lý do cho việc này là “sự bùng nổ tổ hợp” của các khả năng cần phải được khám phá bằng các phương pháp dựa trên các thao tác như tìm kiếm vét cạn (exhaustive search). Các phương pháp này phù hợp với những bài toán đơn giản, nhưng thất bại khi tình huống trở nên phức tạp hơn. Ví dụ, để chứng minh một định lý có phép chứng minh dài năm dòng trong một hệ thống diễn dịch với một quy tắc suy luận và năm tiên đề, chỉ cần liệt kê 3.125 tổ hợp có thể có và kiểm tra từng tổ hợp xem nó có cho kết luận mong muốn hay không. Tìm kiếm vét cạn vẫn có thể dùng được cho các chứng minh sáu hoặc bảy dòng, nhưng khi nhiệm vụ trở nên khó khăn hơn, phương pháp này sẽ gặp rắc rối. Chứng minh một định lý với phép chứng minh dài 50 dòng không chỉ cần nhiều thời gian gấp 10 lần chứng minh định lý có 5 dòng, mà nếu sử dụng tìm kiếm vét cạn, ta sẽ phải rà soát $550 \approx 8,9 \times 10^{34}$ chuỗi khả thể. Đây là một công việc bất khả về mặt tính toán, ngay cả với các siêu máy tính nhanh nhất.

Để vượt qua được sự bùng nổ tổ hợp, người ta cần các thuật toán khai thác được cấu trúc của miền đích và tận dụng kiến thức trước đó bằng phương pháp tìm kiếm tự nghiệm (heuristic), lập kế hoạch và các phép biểu diễn trừu tượng linh hoạt – những năng lực rất kém phát triển trong các hệ thống AI ban đầu. Hiệu năng của những hệ thống ban đầu này cũng bị ảnh hưởng bởi các phương

pháp tậ hại dùng để xử lý bất ổn, sự dựa dẫm vào những phép biểu diễn bằng biểu tượng cứng nhắc và không có cơ sở, sự thiếu hụt dữ liệu và hạn chế quá lớn của phần cứng về dung lượng bộ nhớ và tốc độ bộ xử lý. Tới giữa những năm 1970, nhận thức về những vấn đề này đã tăng lên. Việc nhận ra rằng nhiều dự án AI có thể không bao giờ thực hiện được những lời hứa ban đầu của mình đã dẫn đến “mùa đông AI” đầu tiên: một giai đoạn thất vọng buộc bụng, khi mà ngân sách giảm, sự nghi hoặc tăng lên, và AI trở nên lỗi thời.

Một mùa xuân mới lại đến vào đầu những năm 1980, khi Nhật Bản khởi động Fifth-Generation Computer Systems Project (Dự án Hệ thống Máy tính Thế hệ thứ Năm), một dự án hợp tác công-tư với ngân sách lớn nhằm đi tắt đón đầu công nghệ cao bằng cách phát triển một kiến trúc tính toán song song hàng loạt có thể làm nền tảng cho trí tuệ nhân tạo. Điều này xảy ra một cách đầy màu nhiệm cùng với “phép màu kinh tế hậu chiến” Nhật Bản, giai đoạn mà các chính phủ và các chủ doanh nghiệp phương Tây đang lo lắng kiếm tìm công thức tạo nên sự thành công về kinh tế của Nhật Bản với hy vọng lặp lại phép màu này ở nước mình. Khi Nhật quyết định đầu tư lớn vào AI, một vài quốc gia khác đã lần lượt theo chân.

Những năm tháng sau đó chứng kiến sự sinh sôi nảy nở nhanh chóng của các hệ chuyên gia (expert systems). Được thiết kế như các công cụ hỗ trợ cho người ra quyết định, chúng là những chương trình hoạt động theo quy tắc thực hiện những suy luận đơn giản dựa trên cơ sở kiến thức cung cấp bởi các chuyên gia thuộc nhiều lĩnh vực và được mã hóa thủ công bằng một ngôn ngữ hình thức. Hàng trăm hệ chuyên gia như thế đã được chế tạo. Tuy nhiên, những hệ thống nhỏ chỉ đem lại lợi ích không đáng kể, còn những hệ thống lớn

hơn cho thấy việc phát triển, kiểm định và cập nhật chúng khá tốn kém, và về cơ bản, chúng tương đối cồng kềnh khi sử dụng. Sẽ thật thiếu thực tiễn nếu mua cả một chiếc máy tính chỉ để chạy một chương trình nào đó. Mùa tăng trưởng này cuối cùng cũng kết thúc vào cuối những năm 1980.

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

Tuy nhiên, việc nghiên cứu kỹ thuật đã tiếp diễn nhanh chóng, và đến những năm 1990, mùa đông AI thứ hai đã dần ấm lên. Sự lạc quan được nhóm lên nhờ những kỹ thuật mới, có vẻ như đem lại các phương án thay thế cho mô hình logic truyền thống (thường được gọi là “Good Old-Fashioned Artificial Intelligence – Trí tuệ Nhân tạo Thời xưa”, hay “GOFAI”), vốn tập trung vào việc vận dụng ký hiệu cấp độ cao và đã đạt đến đỉnh cao trong các hệ chuyên gia của những năm 1980. Những kỹ thuật mới được phổ biến, bao gồm mạng lưới thần kinh và các thuật toán di truyền, hứa hẹn bù đắp được một số thiếu hụt của cách tiếp cận GOFAI, cụ thể là “tính dễ đổ vỡ” (brittleness) đặc trưng của các chương trình AI (thường tạo ra những thứ hoàn toàn vô nghĩa nếu các lập trình viên chỉ cần có một giả định

hơi sai lệch). Những kỹ thuật mới có phương pháp vận hành mang tính tổ chức hơn. Ví dụ, các mạng lưới thần kinh có thuộc tính “suy giảm thanh nhã” (hay thoái hóa không đáng kể – graceful degradation): một hư hại nhỏ của mạng lưới thần kinh thường chỉ gây ra một suy giảm nhỏ về hiệu năng của nó thay vì làm sập hoàn toàn. Quan trọng hơn nữa, các mạng lưới thần kinh có thể học hỏi từ kinh nghiệm, tìm kiếm những cách thức tự nhiên để khái quát hóa từ các ví dụ và tìm ra những quy luật thống kê ẩn giấu trong dữ liệu đầu vào.²³ Điều này khiến chúng nhận ra quy luật và phân loại vấn đề một cách hiệu quả. Ví dụ, thông qua đào tạo bằng một tập dữ liệu tín hiệu siêu âm, một mạng lưới thần kinh có thể được dạy để phân biệt các đặc điểm âm thanh của tàu ngầm, mìn và động thực vật biển tốt hơn các chuyên gia – và điều đó có thể thực hiện được mà không cần có người tìm ra trước cách thức xác định các phạm trù phân loại và đánh giá trọng số của các đặc tính khác nhau.

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

Những phẩm chất tương tự não bộ của mạng lưới thần kinh vượt trội hơn so với các hệ thống GOFAL truyền thống dựa trên quy tắc

vốn có phương pháp hoạt động logic cứng nhắc nhưng dễ đổ vỡ. Nó đủ tốt để truyền cảm hứng cho một trường phái mới, kết nối luận (connectionism), nhấn mạnh vào tầm quan trọng của phương pháp xử lý biểu tượng thứ cấp (sub-symbolic) song song. Hơn 150.000 bài báo khoa học đã được công bố về mạng lưới thần kinh nhân tạo, và đây tiếp tục là cách tiếp cận quan trọng trong học máy (machine learning).

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

khi đưa ra một định dạng biểu diễn tối ưu. Nếu không có một cách thức hiệu quả để mã hóa các giải pháp ứng viên (một ngôn ngữ di truyền khớp với cấu trúc tiềm tàng trong miền mục tiêu), việc tìm kiếm tiến hóa thường vòng vèo bất tận trong một không gian tìm kiếm khổng lồ hoặc bị kẹt với một giá trị tối ưu cục bộ. Thậm chí nếu tìm được một định dạng biểu diễn tốt, tiến hóa cũng đòi hỏi nhiều công suất tính toán và thường bị đánh bại bởi sự bùng nổ tổ hợp.

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

gian giải pháp này có các điểm mạnh, điểm yếu riêng có thể được nghiên cứu bằng toán học. Các thuật toán khác nhau về yêu cầu thời gian của bộ xử lý và không gian bộ nhớ với những thiên kiến quy nạp (inductive bias), về sự dễ dàng trong việc tích hợp nội dung được tạo ra từ bên ngoài, và về sự khả kiến trong hoạt động nội bộ của chúng đối với một người phân tích.

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

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

Khung 1: Tác tử Bayes tối ưu

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

Khi tác tử nhận thông tin mới từ các cảm biến của mình, nó cập nhật phân bố xác suất bằng cách điều chỉnh lại phân bố này theo thông tin mới theo định lý Bayes.³⁰ Điều kiện hóa (conditionalization) là phép toán thiết lập xác suất mới của những thế giới không nhất quán với thông tin nhận được để xóa và tái chuẩn hóa phân bố xác suất của những thế giới khả thể còn lại. Kết quả của phép toán này là một “phân bố xác suất hậu nghiệm” (mà tác tử có thể dùng như tiên nghiệm của nó trong bước tiếp theo). Khi tác tử quan sát, khối lượng xác suất của nó tập trung vào một tập đang thu hẹp lại của các thế giới khả thể vẫn còn phù hợp với bằng chứng; và trong tập hợp đó, những thế giới đơn giản hơn luôn có xác suất lớn hơn.

Để hình dung, ta có thể so sánh xác suất với cát trên một tờ giấy lớn. Tờ giấy được chia thành các vùng có kích thước khác nhau, mỗi vùng tương ứng với một thế giới khả thể, vùng có diện tích lớn hơn tương ứng với những thế giới đơn giản hơn. Hãy hình dung thêm một lớp cát có độ dày đồng đều được rải lên cả tờ giấy: đó chính là phân bố xác suất tiên nghiệm của chúng ta. Mỗi khi một quan sát

được thực hiện và loại bỏ một số thế giới, chúng ta gạt cát ra khỏi vùng tương ứng của tờ giấy và phân phối lại cát đồng đều trên những vùng còn lại. Như vậy tổng lượng cát trên tờ giấy không thay đổi, mà chỉ tập trung vào ít vùng hơn khi bằng chứng quan sát được tích lũy. Đây là bức tranh về quá trình học ở dạng thức thuần khiết nhất. (Để tính toán xác suất của một giả thuyết, ta chỉ cần đo lượng cát ở tất cả các vùng tương ứng với những thế giới khả thể mà trong đó giả thuyết là đúng).

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

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

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

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

Một ưu điểm của việc liên kết vấn đề học tập từ những miền cụ thể với vấn đề chung của suy luận Bayes là những thuật toán mới làm cho suy luận Bayes hiệu quả hơn sau đó sẽ tạo nên những cải thiện tức thời trong nhiều lĩnh vực khác nhau. Ví dụ, những tiến bộ trong các kỹ thuật tính xấp xỉ của Monte Carlo đang được áp dụng trực tiếp trong thị giác máy, robotic và di truyền học điện toán. Một ưu

điểm khác là nó cho phép các nhà nghiên cứu từ các lĩnh vực khác nhau hợp nhất kết quả nghiên cứu của mình dễ dàng hơn. Những mô hình đồ họa và thống kê Bayes trở thành tiêu điểm nghiên cứu chung trong nhiều lĩnh vực, bao gồm học máy, vật lý thống kê, tin học sinh học, tối ưu hóa tổ hợp và lý thuyết truyền thông.³⁵ Những tiến bộ đáng kể gần đây trong học máy có được từ việc tổ hợp các kết quả chính thống khởi phát trong các lĩnh vực học thuật khác. (Các ứng dụng học máy cũng có được lợi ích rất lớn từ những máy tính nhanh hơn và tính sẵn sàng cao hơn của các tập dữ liệu lớn.)

Những công nghệ tiên tiến

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

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

Tuy nhiên, trong việc này còn có một ý nghĩa quan trọng khác. Đó là việc AI chơi cờ vua hóa ra lại là một “chiến thắng” khiêm tốn so với những gì nhiều người vẫn tưởng. Người ta từng giả định, có thể

cũng không hoàn toàn vô căn cứ, rằng để máy tính có thể chơi được cờ vua ở cấp đại kiện tướng, nó phải được trang bị trí tuệ tổng quát ở cấp độ cao.³⁹ Ví dụ, người ta có thể nghĩ rằng để chơi cờ giỏi cần có khả năng học những khái niệm trừu tượng, suy nghĩ thông minh về chiến lược, lập những kế hoạch linh hoạt, thực hiện nhiều phép suy luận logic độc đáo và thậm chí có thể là lập mô hình phương thức tư duy của đối thủ. Tuy nhiên mọi thứ không phải như vậy. Ta có thể xây dựng một công cụ chơi cờ vua giỏi hoàn hảo từ một thuật toán có mục đích riêng biệt.⁴⁰ Khi được thực thi trên các bộ xử lý nhanh xuất hiện trong những năm cuối thế kỷ ²⁰, nó đã tạo ra một năng lực chơi cờ mạnh mẽ. Nhưng một AI được chế tạo như vậy lại chỉ có khả năng hạn hẹp: Nó biết chơi cờ nhưng không làm được gì khác.⁴¹

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

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

Trong những lĩnh vực khác, các giải pháp hóa ra lại phức tạp hơn so với kỳ vọng ban đầu và cũng tiến triển chậm hơn. Nhà khoa học máy tính Donald Knuth đã ngạc nhiên với việc “AI cho tới nay đã thành công khi làm hầu hết mọi thứ đòi hỏi ‘tư duy’ nhưng không làm được hầu hết những thứ mà người và động vật có thể làm mà ‘không cần suy nghĩ’ – điều này, bằng cách nào đó, hóa ra lại khó hơn nhiều!”⁶⁰ Việc phân tích những khung cảnh trực quan, nhận diện đối tượng hoặc kiểm soát hành vi của một robot khi tương tác với môi trường tự nhiên tỏ ra là những việc vô cùng thách thức. Mặc dù vậy, chúng ta đã và đang đạt được nhiều tiến bộ, và những tiến bộ này được hỗ trợ bởi những cải thiện phần cứng đáng kể.

Hiểu lẽ thường và hiểu ngôn ngữ tự nhiên cũng là những vấn đề khó khăn. Giờ đây, người ta thường nghĩ rằng việc đạt trình độ hoàn toàn ngang với con người khi thực thi những nhiệm vụ này là vấn đề “AI toàn diện” (AI-complete), có nghĩa là độ khó của việc giải quyết

những vấn đề này về cơ bản tương đương với độ khó của việc chế tạo ra những chiếc máy có trí tuệ tổng quát cấp độ con người.⁶¹ Nói cách khác, nếu ai đó thành công trong việc tạo ra một AI có thể hiểu ngôn ngữ tự nhiên tốt như một người lớn thì họ hoàn toàn có thể tạo ra một AI làm được mọi việc khác mà con người có thể, hoặc họ đã đến được rất gần với một năng lực tổng thể như vậy.⁶²

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

Các phương pháp trí tuệ nhân tạo hiện đang được sử dụng trong quá nhiều lĩnh vực để có thể xem xét tất cả ở đây, nhưng một ví dụ sẽ cho ta ý niệm về sự trải rộng của các ứng dụng. Ngoài những AI

chơi trò chơi được liệt kê trong Bảng 1 còn có những thiết bị trợ thính với các thuật toán lọc tiếng ồn xung quanh; những bộ tìm đường hiển thị bản đồ và chỉ đường cho lái xe; các hệ thống đề xuất sách hoặc album nhạc trên cơ sở những lần mua và đánh giá trước đó của một người dùng; và những hệ thống ra quyết định y tế giúp các bác sĩ chẩn đoán ung thư vú, khuyến nghị phác đồ điều trị và hỗ trợ đọc điện tâm đồ. Chúng ta có những robot thú cưng, robot vệ sinh, robot cắt cỏ, robot giải cứu, robot phẫu thuật và hơn một triệu robot công nghiệp.⁶⁴ Số lượng robot thế giới đã vượt quá 10 triệu con.⁶⁵

Nhận dạng giọng nói hiện đại dựa trên những kỹ thuật thống kê như các mô hình Markov ẩn đã đủ chính xác để áp dụng trong thực tiễn (một số nội dung trong cuốn sách này được soạn thảo với sự trợ giúp của một chương trình nhận dạng giọng nói). Các trợ lý cá nhân kỹ thuật số, như Siri của Apple, phản hồi với mệnh lệnh bằng giọng nói và có thể trả lời những câu hỏi đơn giản và thực hiện các lệnh. Kỹ thuật nhận dạng ký tự quang học đối với các văn bản viết tay và đánh máy thường xuyên được sử dụng trong các ứng dụng như phân loại thư tín và số hóa các tài liệu cũ.⁶⁶

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

lập trình viên xây dựng những hệ thống này thậm chí không nói thứ ngôn ngữ mà họ đang xử lý.⁶⁷

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

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

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

Lập thời gian biểu thông minh là một trong những lĩnh vực thành công quan trọng. Công cụ lập kế hoạch và thời gian biểu tự động DART phục vụ hậu cần được sử dụng trong Chiến dịch Bão sa mạc

năm 1991 hiệu quả tới mức DARPA (Cơ quan Dự án nghiên cứu quốc phòng tiên tiến của Mỹ) tuyên bố rằng chỉ ứng dụng duy nhất này đã đủ hoàn vốn toàn bộ đầu tư trong suốt 30 năm của họ vào AI.⁶⁸

Các hệ thống đặt vé máy bay sử dụng các hệ thống lập thời gian biểu và định giá phức tạp. Các doanh nghiệp cũng ứng dụng rộng rãi kỹ thuật AI vào các hệ thống kiểm soát kho hàng. Họ cũng dùng các hệ thống đặt hàng tự động và trợ giúp qua điện thoại, được kết nối với phần mềm nhận dạng giọng nói để hướng dẫn những khách hàng kém may mắn vượt qua một mê lộ các menu đan lồng phức tạp.

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

Giờ đây, ta phải nhấn mạnh rằng ranh giới giữa trí tuệ nhân tạo và phần mềm nói chung đã không còn rõ nét. Một số ứng dụng được liệt kê ở trên có thể được xem như các ứng dụng phần mềm chung thay vì AI cụ thể, mặc dù điều đó đưa chúng ta quay lại với tuyên bố của McCarthy, rằng khi thứ gì đó làm được việc thì nó sẽ không còn là AI nữa. Một cách phân biệt khác phù hợp hơn với mục tiêu của

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

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

tới hơn một nửa lượng cổ phiếu được giao dịch trên thị trường Mỹ.⁶⁹ Hoạt động giao dịch thuật toán cũng có dính líu tới vụ sụp đổ chớp nhoáng thị trường năm 2010 (2010 Flash Crash) (xem Khung 2).

Khung 2: Vụ Flash Crash năm 2010

Đến chiều ngày 6/5/2010, thị trường chứng khoán Mỹ đã giảm 4% do những lo lắng về khủng hoảng nợ châu Âu. Vào 2h32 chiều, một người bán lớn (một tổ hợp quỹ tương hỗ) đã khởi động một thuật toán bán để bán đi một số lượng lớn hợp đồng tương lai E-Mini S&P 500 với giá bán được liên kết với một chỉ số đo thanh khoản từng phút trên thị trường chứng khoán. Những hợp đồng này được mua thông qua chương trình giao dịch thuật toán cao tần, được lập trình để nhanh chóng loại bỏ thế giá lên (long position) tạm thời bằng cách bán các hợp đồng cho các chương trình giao dịch thuật toán khác. Khi nhu cầu từ của những người mua cơ bản suy yếu, trình giao dịch thuật toán bắt đầu bán E-Mini cho các trình giao dịch thuật toán khác (sơ cấp), và các trình này lại tiếp tục bán cho các trình giao dịch thuật toán khác nữa (thứ cấp), tạo ra một hiệu ứng “củ khoai nóng” (hot potato) đẩy khối lượng giao dịch lên cao – điều đó được thuật toán bán diễn dịch như một chỉ số cho thấy tính thanh khoản cao, thúc đẩy nó tăng tốc độ đưa các hợp đồng E-Mini ra thị trường, tạo ra một xoáy ốc đi xuống. Đến một thời điểm nào đó, các chương trình giao dịch thuật toán cao tần bắt đầu rút khỏi thị trường, làm giảm tính thanh khoản trong lúc giá tiếp tục giảm. Đến 2h45, giao dịch E-Mini bị tạm dừng bởi một bộ ngắt mạch tự động, chức năng chặn logic (stop logic functionality) của sàn giao dịch. Khi giao dịch được khôi phục, chỉ năm giây sau đó, giá ổn định lại và sau đó bắt đầu bù đắp được phần lớn thiệt hại. Nhưng trong một khoảng thời gian, dưới đáy của

cuộc khủng hoảng, thị trường mất một ngàn tỷ đô la, và hiệu ứng “tràn” đã dẫn đến một số lượng đáng kể giao dịch chứng khoán riêng lẻ được thực hiện với mức giá “lố bịch”, như 1 cent hay 100.000 đô la. Sau khi thị trường đóng cửa giao dịch ngày hôm đó, đại diện của các sàn giao dịch đã gặp các nhà quản lý và quyết định hủy mọi giao dịch được thực hiện với mức giá chênh lệch 60% hoặc cao hơn so với mức trước khủng hoảng (coi những giao dịch này là “sai lỗi rõ ràng” và do đó sẽ bị huỷ bỏ hồi tố (post facto cancellation) theo các quy tắc giao dịch hiện hành.⁷⁰

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

Một bài học khác là các chuyên gia thông minh có thể ra lệnh cho một chương trình trên cơ sở một giả định có vẻ có nghĩa và thông thường là đúng (ví dụ, khối lượng giao dịch là một số đo tốt của tính thanh khoản thị trường), và điều đó có thể tạo ra những kết quả thảm họa khi chương trình tiếp tục tuân lệnh với sự nhất quán logic cứng nhắc ngay cả trong tình huống bất thường khi giả định không còn

đúng. Thuật toán chỉ làm những gì nó thường làm; và trừ khi có một loại thuật toán rất đặc biệt, nó sẽ không quan tâm chúng ta có ôm đầu tuyệt vọng hay hỗn hển vì hoảng sợ vì sự không thích hợp tới mức lối bịch của những hành động của nó. Đó là một chủ đề mà chúng ta sẽ còn gặp lại.

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

Những quan điểm về tương lai của trí tuệ máy

Tiến bộ trên hai mặt trận chính – một mặt hướng tới một nền tảng thống kê và lý thuyết thông tin vững chắc hơn cho học máy, và mặt khác hướng tới thành công thực tiễn và thương mại của các ứng dụng đặc thù ứng dụng hoặc đặc thù miền – đã trả lại cho nghiên cứu AI một số đặc quyền đã mất. Tuy nhiên, hiệu ứng văn hóa của lịch sử trước đó lên cộng đồng AI có thể vẫn còn tồn dư và khiến nhiều nhà nghiên cứu chính thống không muốn điều chỉnh lại mình theo những tham vọng quá mức. Do đó Nils Nilsson, một trong những người kỳ cựu trong lĩnh vực này, than phiền rằng những đồng nghiệp hiện nay thiếu sự dũng cảm về tinh thần đã từng thúc đẩy thế hệ của ông:

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

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

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

Ý kiến của các chuyên gia về tương lai của AI rất đa chiều. Có một sự không đồng thuận về thang thời gian cũng như dạng thức cuối cùng của AI. Theo một nghiên cứu mới đây, những dự báo về

phát triển tương lai của trí tuệ nhân tạo “đa dạng đến đâu thì đáng tin cậy đến đó”.⁷⁶

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

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

Các kết quả khảo sát cũng tương đồng với một số cuộc phỏng vấn mới được công bố gần đây với khoảng hơn 20 nhà nghiên cứu trong các lĩnh vực liên quan đến AI. Chẳng hạn, Nils Nilsson đã có một sự nghiệp lâu dài và hiệu quả dành cho những vấn đề về tìm kiếm, lập kế hoạch, biểu diễn tri thức và người máy. Ông đã viết sách giáo khoa về trí tuệ nhân tạo, và gần đây đã hoàn thành bộ sử toàn

diện nhất từng được viết cho tới nay về lĩnh vực này.⁷⁹ Khi được hỏi về ngày “đổ bộ” của HLMI, ông đưa ra những ý kiến sau:⁸⁰

10% khả năng: 2030

50% khả năng: 2050

90% khả năng: 2100

Bảng 2 Khi nào chúng ta có được trí tuệ máy cấp độ con người?⁸¹			
	10 %	50 %	90 %
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Tổng hợp	2022	2040	2075

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

Theo quan điểm của riêng tôi, những con số trung bình đưa ra trong khảo sát chuyên gia chưa có đủ khối lượng xác suất về thời

gian xuất hiện muộn hơn. Xác suất 10% HLMI chưa được phát triển vào những năm 2075 hoặc thậm chí 2100 (sau khi đã điều chỉnh theo điều kiện “hoạt động khoa học của loài người tiếp tục diễn ra mà không gặp phải gián đoạn tiêu cực lớn”) xem ra quá thấp.

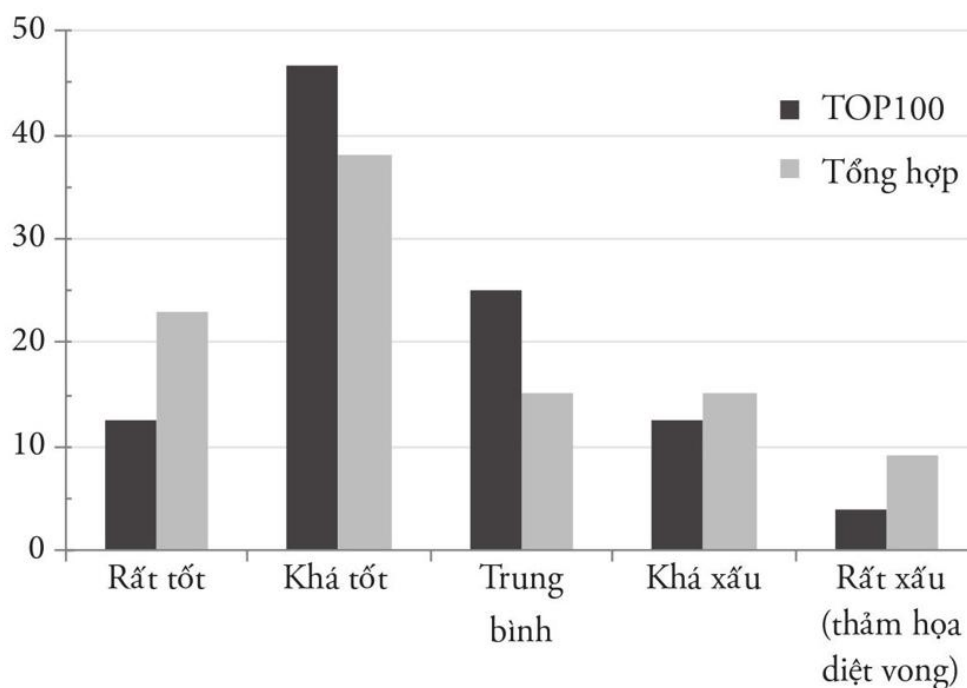
Về mặt lịch sử, các nhà nghiên cứu AI chưa từng có thành tựu nổi bật về khả năng dự đoán tốc độ phát triển trong chính lĩnh vực của họ hay bóng dáng tương lai của những phát triển đó. Một mặt, một số nhiệm vụ như chơi cờ vua hóa ra lại có thể thực hiện được bằng những chương trình đơn giản đến ngạc nhiên; và những kẻ ác khẩu từng nói rằng máy móc sẽ “không bao giờ” làm được việc này hay việc kia đã thường xuyên bị chứng minh là sai lầm. Mặt khác, những lỗi lầm tiêu biểu hơn của giới chuyên môn cho thấy khó khăn trong việc làm cho một hệ thống thực hiện được một cách chắc chắn những nhiệm vụ thực tế vẫn đang bị đánh giá thấp, và những tiến bộ của một vài dự án hay kỹ thuật ưa thích cụ thể của họ đang được đánh giá quá cao.

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

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

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

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



Hình 2: Tác động tổng thể lâu dài của HLMI.⁸³

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

Kích thước mẫu bé, thiên kiến lựa chọn và trên hết là sự không đáng tin cậy cố hữu của các quan điểm chủ quan đồng nghĩa với

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

Chương 2

Con đường đến với siêu trí tuệ

Máy móc hiện tại vẫn yếu hơn rất nhiều so với con người về mặt trí tuệ tổng quát. Nhưng một ngày nào đó (như chúng tôi đã nói), chúng sẽ trở thành siêu trí tuệ. Vậy lối đi nào dẫn chúng ta đến đó? Chương này sẽ khám phá một số con đường công nghệ để hình dung nhất. Chúng ta sẽ xem xét vấn đề trí tuệ nhân tạo, giả lập hoàn chỉnh não bộ (whole brain emulation), nhận thức sinh học và giao diện người-máy, cũng như mạng lưới và các tổ chức. Chúng ta cũng sẽ đánh giá những cấp độ khác nhau của tính hợp lý như những con đường dẫn đến siêu trí tuệ. Sự tồn tại của nhiều con đường sẽ làm tăng xác suất tới đích của chúng ta khi đi qua ít nhất một trong số chúng.

Chúng ta có thể tạm định nghĩa siêu trí tuệ là bất kỳ trí tuệ nào có thể vượt xa năng lực nhận thức của con người trong hầu hết mọi lĩnh vực cần quan tâm.¹ Chúng tôi sẽ nói nhiều hơn về khái niệm siêu trí tuệ trong chương sau. Khi đó, khái niệm này sẽ được phân tích phổ (spectral analysis) để phân biệt một số dạng thức siêu trí tuệ có thể có. Nhưng hiện tại, chúng ta chỉ cần mô tả nó sơ lược như đã nói. Hãy lưu ý rằng định nghĩa này không đảm bảo về cách thực thi siêu trí tuệ. Nó cũng không phải là một sự cam kết về trải nghiệm chủ quan (qualia): việc siêu trí tuệ có trải nghiệm ý thức chủ quan hay

không có thể đóng vai trò rất quan trọng trong một số vấn đề (đặc biệt là đạo đức), nhưng điều mà chúng tôi quan tâm nhất là các tiền đề và hệ quả nhân quả của siêu trí tuệ, chứ không phải các đặc tính siêu hình học của tinh thần.²

Chương trình Deep Fritz không phải là một siêu trí tuệ theo định nghĩa này vì trí tuệ của nó chỉ giới hạn trong môn cờ vua. Tuy nhiên, một số loại hình siêu trí tuệ có tính đặc thù về mặt lĩnh vực cũng có thể trở nên quan trọng. Khi nói về hiệu năng siêu trí tuệ giới hạn trong một lĩnh vực cụ thể, chúng ta sẽ đưa ra lưu ý cụ thể về giới hạn đó. Ví dụ, một “siêu trí tuệ kỹ thuật” có thể làm việc tốt hơn rất nhiều so với những bộ óc giỏi nhất hiện nay trong lĩnh vực kỹ thuật. Trừ các trường hợp được lưu ý, nhìn chung, chúng ta sử dụng thuật ngữ này để nói về các hệ thống có trí tuệ tổng quát ở cấp độ siêu nhân.

Nhưng chúng ta có thể tạo ra siêu trí tuệ như thế nào? Hãy cùng nghiên cứu một số con đường khả thi.

Trí tuệ nhân tạo

Những người đọc chương này không nên mong đợi một bản thiết kế cho việc lập trình một trí tuệ nhân tạo tổng quát. Tất nhiên vẫn chưa có một bản thiết kế nào như vậy tồn tại. Ngay cả nếu tôi có bản thiết kế đó đi nữa, chắc chắn tôi sẽ không công bố nó trong một cuốn sách. (Nếu lý do cho việc này chưa lập tức trở nên hiển hiện thì những luận điểm trong các chương sau sẽ làm rõ nó hơn.)

Tuy nhiên, chúng tôi có thể nhận biết được một số tính năng chung của loại hình hệ thống cần có. Hiện có vẻ khá rõ ràng rằng năng lực học tập sẽ là một tính năng tích hợp trong thiết kế lõi của một hệ thống được định hướng để có được trí tuệ tổng quát, không phải là thứ có thể xử lý sau như một phần mở rộng hoặc bổ sung.

Khả năng xử lý hiệu quả tính bất định và thông tin xác suất cũng tương tự. Một số năng lực dùng trong việc trích xuất những khái niệm hữu ích từ dữ liệu giác quan và các trạng thái bên trong; việc nâng cấp những khái niệm thu được thành các phép biểu diễn tổ hợp để sử dụng trong suy luận logic và trực giác cũng có thể sẽ nằm trong những tính năng thiết kế lõi của một AI hiện đại được định hướng để có được trí tuệ tổng quát.

Những hệ thống Trí tuệ Nhân tạo Thời xưa hầu hết không tập trung vào khả năng học tập, tính bất định hay tạo dựng khái niệm, có thể vì các kỹ thuật dùng để xử lý những vấn đề này khi đó vẫn còn kém phát triển. Điều này không đồng nghĩa với việc những ý tưởng nền tảng là hoàn toàn lạ lẫm. Ý tưởng sử dụng học tập như một công cụ đưa một hệ thống đơn giản hơn lên trình độ trí tuệ cấp độ con người ít nhất cũng bắt nguồn từ ý niệm của Alan Turing về một “cỗ máy trẻ con” (child machine), mà ông viết năm 1950:

Thay vì cố gắng tạo ra một chương trình mô phỏng trí óc người lớn, sao chúng ta không thử tạo ra một chương trình mô phỏng trí óc trẻ em? Nếu sau đó được “giáo dục” phù hợp, nó có thể có được não bộ của người lớn.³

Turing đã mừng tượng ra một quá trình lặp đi lặp lại nhằm phát triển một cỗ máy trẻ con như vậy:

Chúng ta không thể kỳ vọng tìm được một cỗ máy trẻ con giỏi giang ngay từ đầu mà phải thử nghiệm bằng việc dạy một cỗ máy như vậy và xem nó học tốt đến đâu. Sau đó, ta phải thử với một cỗ máy khác để xem liệu nó tốt hay kém hơn. Có một sự liên kết rõ ràng giữa quá trình này và sự tiến hóa... Tuy nhiên, ta có thể hy vọng rằng quá trình này sẽ diễn ra nhanh hơn. Sự sống sót của cá thể phù hợp

nhất là một phương pháp chậm chạp để đo lường tiến bộ. Người tiến hành thử nghiệm, bằng cách vận dụng trí tuệ, sẽ đẩy nhanh được quá trình. Anh ta không bị giới hạn bởi biến dị ngẫu nhiên, và điều này cũng quan trọng không kém. Nếu người này có thể truy tìm nguyên nhân của một điểm yếu nào đó, thì sẽ nghĩ ra được kiểu biến dị để cải thiện nó.⁴

Chúng ta biết rằng các quá trình tiến hóa mù quáng có thể tạo ra trí tuệ tổng quát cấp độ con người vì ít nhất chúng cũng đã làm được việc đó một lần. Các quá trình tiến hóa có tầm nhìn xa – là các chương trình di truyền được thiết kế và hướng dẫn bởi một lập trình viên thông minh là con người – có thể đạt được kết quả tương tự với hiệu quả cao hơn nhiều. Quan sát này đã được một số triết gia và nhà khoa học, bao gồm David Chalmer và Hans Moravec, sử dụng để lập luận rằng AI cấp độ con người không chỉ có thể về mặt lý thuyết mà còn khả thi ngay trong thế kỷ này.⁵ Ý tưởng là chúng ta có thể ước lượng những năng lực có liên quan của tiến hóa và kỹ thuật của con người trong việc tạo ra trí tuệ, và thấy rằng kỹ thuật con người đã vượt trội hơn nhiều so với tiến hóa trong một số lĩnh vực và chẳng mấy chốc có thể sẽ vượt trội trong những lĩnh vực còn lại. Vì vậy, việc tiến hóa tạo ra trí tuệ cho thấy kỹ thuật của con người cũng sẽ sớm làm được điều tương tự. Vì vậy Moravec đã viết (năm 1976):

Sự tồn tại của một vài ví dụ về trí tuệ được thiết kế trong những hạn chế này sẽ mang lại một niềm tin to lớn rằng chúng ta có thể đạt được kết quả như vậy trong thời gian ngắn hơn. Tình huống này cũng tương tự lịch sử của những chuyến bay “nặng hơn không khí”,

trong đó chim chóc, dơi và côn trùng đã trình diễn khả năng này trước khi con người thành thạo nó.⁶

Tuy nhiên, ta vẫn cần thận trọng về những kết luận rút ra từ hướng suy luận này. Đúng là tiến hóa tạo ra khả năng bay của vật nặng hơn không khí, và kỹ thuật của con người sau đó đã thành công trong việc làm chủ nó (mặc dù bằng một cơ chế rất khác). Chúng ta cũng có thể đưa ra những ví dụ khác như siêu âm, định hướng từ trường, vũ khí hóa học, bộ thu quang học và toàn bộ các loại đặc tính cơ học, vận động. Tuy nhiên, ta cũng có thể chỉ ra những lĩnh vực mà các kỹ sư không theo kịp tiến hóa, chẳng hạn như tạo hình cơ thể, tự sửa chữa và phòng vệ miễn dịch, trong đó những nỗ lực của con người tụt hậu rất xa so với những gì thiên nhiên đã làm được. Vì thế, luận điểm của Moravec không thể cho chúng ta “niềm tin to lớn” rằng mình có thể đạt được trí tuệ nhân tạo cấp độ con người “trong thời gian ngắn”. Khả quan nhất, quá trình tiến hóa của sự sống thông minh đặt ra một mức trần cho khó khăn vốn có của việc thiết kế trí tuệ, nhưng mức trần này có thể cao hơn rất nhiều so với năng lực kỹ thuật hiện có của con người.

Một luận điểm tiến hóa khác về tính khả thi của AI có thể được trình bày thông qua ý tưởng rằng chúng ta có thể đạt được những kết quả tương đương với tiến hóa sinh học bằng cách chạy các thuật toán di truyền trên những máy tính đủ nhanh. Phiên bản này của luận điểm tiến hóa đề xuất một phương pháp cụ thể có thể tạo ra trí tuệ.

Nhưng có đúng là chúng ta sắp có đủ công suất tính toán để tóm tắt lại các quá trình tiến hóa có liên quan đã tạo ra trí tuệ con người? Câu trả lời phụ thuộc cả vào việc công nghệ tính toán sẽ phát triển bao nhiêu trong những thập kỷ tới, và cần công suất tính toán ra sao

để chạy các thuật toán di truyền với công suất tối ưu hóa giống như quá trình tiến hóa của chọn lọc tự nhiên trong quá khứ. Mặc dù cuối cùng, kết luận mà chúng ta có được từ việc theo đuổi đường hướng suy luận này là mập mờ một cách đáng thất vọng nhưng việc nỗ lực tiến hành đánh giá sơ bộ vẫn có ý nghĩa hướng dẫn (xem Khung 3). Hoạt động này ít nhất cũng sẽ thu hút được sự chú ý đến một vài ẩn số thú vị.

Kết luận lại, các nguồn tài nguyên tính toán cần có để sao chép một cách đơn giản các quá trình tiến hóa có liên quan trên Trái đất đã tạo ra trí tuệ cấp độ con người vẫn rất xa tầm với – và sẽ còn tiếp tục như vậy ngay cả khi định luật Moore tiếp tục đúng trong một thế kỷ nữa (Hình 3). Tuy nhiên, so với việc sao chép tổng lực (brute-force) các quá trình tiến hóa tự nhiên, hiệu năng có thể gia tăng đáng kể bằng cách thiết kế quá trình tìm kiếm nhắm đến trí tuệ, sử dụng vô số cải tiến rõ nét trên chính chọn lọc tự nhiên. Mặc dù vậy, rất khó biết chắc được những sự gia tăng hiệu năng có thể có được này là bao nhiêu. Chúng ta thậm chí còn không thể nói được liệu chúng có thể tăng đến năm hay 25 cấp khuếch đại^a (105 đến 1025 lần). Vì không có lời giải thích cận kề, các luận điểm tiến hóa không thể giới hạn được những mong đợi của chúng ta về khó khăn của việc xây dựng trí tuệ máy cấp độ con người, hay về thang thời gian của những phát triển này.

^a *Cấp khuếch đại: Order of magnitude. (BTV)*

Khung 3: Chúng ta cần điều gì để tóm lược quá trình tiến hóa?

Không phải thành công nào mà tiến hóa có được trong quá trình phát triển trí tuệ con người cũng đều liên quan đến việc một kỹ sư cổ

gắng phát triển trí tuệ máy bằng phương pháp nhân tạo. Chỉ một phần nhỏ lựa chọn tiến hóa trên Trái đất là lựa chọn cho trí tuệ. Cụ thể hơn, những vấn đề mà các kỹ sư không thể bình thản cho qua có thể chỉ là mục tiêu của một phần rất nhỏ trong quá trình lựa chọn tự nhiên tổng thể. Ví dụ, khi có thể chạy máy tính bằng điện, chúng ta không cần phát minh lại các phân tử trong ngành năng lượng tế bào để tạo ra máy móc thông minh – mặc dù vậy sự tiến hóa phân tử này của các lộ trình chuyển hóa có thể sử dụng một phần lớn năng lượng lựa chọn dành cho tiến hóa trong toàn bộ tiến trình lịch sử của Trái đất.⁷

Ta có thể lý luận rằng những vấn đề chủ chốt của AI được biểu hiện trong cấu trúc của các hệ thần kinh, và chúng xuất hiện cách đây chưa đến một tỷ năm.⁸ Nếu chúng ta chấp nhận quan điểm này thì số “thí nghiệm” phù hợp đối với tiến hóa sẽ được rút gọn đáng kể. Ngày nay chúng ta có khoảng $4-6 \times 10^{30}$ sinh vật nhân sơ trên thế giới, nhưng chỉ có 1019 con bò sát và chưa đến 1010 con người (trong khi dân số thời kỳ tiền nông nghiệp thấp hơn một cấp khuếch đại).⁹ Những con số này mới chỉ đáng sợ ở mức trung bình.

Tuy nhiên các thuật toán tiến hóa không chỉ đòi hỏi các biến dị để lựa chọn mà còn cần một hàm thích nghi để đánh giá các biến dị, và đó thường là phần tử đắt đỏ nhất về mặt tính toán. Một hàm thích nghi cho sự tiến hóa của trí tuệ nhân tạo có thể yêu cầu mô phỏng quá trình phát triển neuron, học tập và nhận thức để đánh giá tính thích nghi. Do đó, chúng ta không nên nhìn vào số lượng sơ bộ các cá thể với hệ thần kinh phức tạp, mà hãy chú ý đến số lượng neuron trong các cá thể sinh học mà chúng ta có thể phải mô phỏng để bắt chước hàm thích nghi của quá trình tiến hóa. Chúng ta có thể đánh

giá sơ bộ số lượng này bằng cách xem xét côn trùng, loài đang thống trị sinh khối của các động vật trên cạn. (Theo đánh giá, riêng loài kiến đã chiếm 15-20%.)¹⁰ Kích thước bộ não côn trùng tương đối khác nhau. Những loài côn trùng lớn và có tính xã hội sở hữu não bộ lớn hơn: bộ não của một con ong mật chỉ có chưa đến 106 neuron, một con ruồi giấm có bộ não với 105 neuron và kiến nằm giữa với 250.000 neuron.¹¹ Phần lớn các loài côn trùng nhỏ hơn có thể có não bộ chỉ khoảng vài ngàn neuron. Một cách tương đối, nếu chúng ta ấn định rằng tất cả 1019 cá thể côn trùng đều có số neuron bằng với ruồi giấm, thì tổng số neuron của côn trùng trên thế giới là 1024.

Con số này có thể được tăng thêm một cấp khuếch đại khi tính đến cả động vật giáp xác sống dưới nước, chim, bò sát hay động vật có vú, để đạt tới giá trị 1025. (Ngược lại trong thời kỳ tiền nông nghiệp, có chưa đến 107 cá thể người, với chưa đến 1011 neuron một cá thể; như vậy tổng số neuron của loài người là chưa đến 1018, mặc dù con người có số synapse (khớp thần kinh) trên một neuron cao hơn).

Chi phí tính toán để mô phỏng một neuron tùy thuộc vào mức độ chi tiết được đưa vào mô phỏng. Những mô hình neuron cực kỳ đơn giản sử dụng khoảng 1.000 phép tính dấu chấm động mỗi giây (FLOPS) để mô phỏng một neuron (theo thời gian thực). Mô hình điện sinh lý thực Hodgkin-Huxley sử dụng 1.200.000 FLOPS. Một mô hình đa khoang vùng (multi-compartmental) chi tiết hơn có thể tăng thêm ba hay bốn cấp khuếch đại nữa, trong khi những mô hình cấp cao hơn trừu tượng hóa được các hệ thống neuron có thể giảm bớt hai tới ba cấp khuếch đại từ các mô hình đơn giản.¹² Nếu mô phỏng 1025 neuron trong suốt một tỷ năm tiến hóa (dài hơn sự tồn tại của

các hệ thống thần kinh như chúng ta biết), và cho phép các máy tính của mình chạy trong một năm, yêu cầu tính toán sẽ rơi vào khoảng từ 1031 – 1044 FLOPS. Để so sánh, máy tính Thiên Hà 2 của Trung Quốc, siêu máy tính mạnh nhất thế giới tính đến tháng 9 năm 2013, cũng chỉ đưa ra được yêu cầu tính toán vào khoảng $3,39 \times 10^{16}$ FLOPS. Trong những thập kỷ gần đây, phải mất khoảng 6,7 năm để các máy tính phổ thông tăng công suất lên một cấp khuếch đại. Ngay cả khi định luật Moore luôn đúng trong suốt một thế kỷ cũng không đủ để lấp đầy khoảng cách, và việc chạy những phần mềm chuyên dụng hơn, hoặc cho phép thời gian chạy lâu hơn, cũng chỉ làm tăng thêm một vài cấp khuếch đại mà thôi.

Xét trên một góc độ khác, đây cũng là một con số khá dè dặt. Quá trình tiến hóa tạo ra trí tuệ con người nhưng lại không “nhắm” tới kết quả này. Nói cách khác, các hàm thích nghi cho các cá thể tự nhiên không chỉ lựa chọn trí tuệ và các dạng tiền thân của nó.¹³ Thậm chí những môi trường mà trong đó, các cá thể có kỹ năng xử lý thông tin vượt trội sẽ gặt hái được các phần thưởng khác nhau có thể cũng không lựa chọn trí tuệ, vì những cải thiện về trí tuệ có thể (và thường xuyên) đòi hỏi các “khoản chi” đáng kể, như mức độ tiêu thụ năng lượng cao hơn hay thời gian trưởng thành lâu hơn. Những chi phí này có thể cao hơn bất kỳ lợi ích nào nhận được từ hành vi thông minh hơn. Những môi trường quá “chết chóc” cũng làm giảm giá trị của trí tuệ: tuổi thọ kỳ vọng của một cá thể càng ngắn thì khả năng học tập được cải thiện càng có ít thời gian để bù đắp lại.

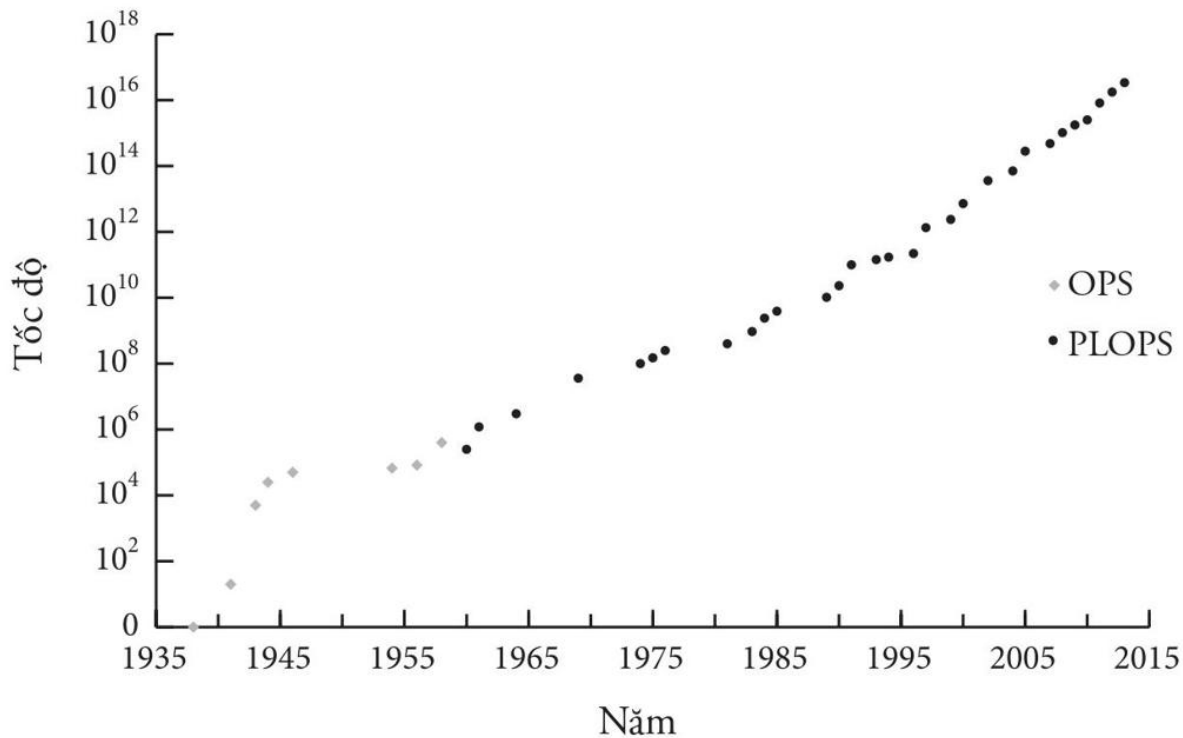
Áp lực chọn lọc được giảm nhẹ đối với trí tuệ làm chậm việc lan tỏa của những sáng tạo làm tăng cường trí tuệ, và như vậy cũng sẽ kìm hãm cơ hội của quá trình chọn lọc để thiết lập ưu tiên với những

sáng tạo tiếp theo, vốn phụ thuộc vào chúng. Hơn nữa, tiến hóa có thể bị mắc kẹt trong những trường hợp tối ưu cục bộ mà con người có thể nhận ra và vượt qua được bằng cách điều chỉnh những sự đánh đổi giữa khai thác và khám phá, hoặc bằng cách đưa ra một tiến trình thử nghiệm mượt mà với những phép thử về trí tuệ có độ khó tăng dần.¹⁴ Như đã nhắc đến ở trên, quá trình tiến hóa phân bố “quyền năng” chọn lọc của mình vào các tính trạng không liên quan đến trí tuệ (như lý thuyết về những cuộc đua đồng tiến hóa cạnh tranh giữa hệ thống miễn dịch và các loài ký sinh của Red Queen). Tiến hóa tiếp tục lãng phí nguồn tài nguyên vào việc tạo ra những đột biến chết người, và cũng không thể tận dụng được những tương đồng thống kê về mặt tác động của các loại đột biến khác nhau. Đây là những điểm kém hiệu năng trong quá trình chọn lọc tự nhiên (khi được xem xét như công cụ tiến hóa trí tuệ). Tuy nhiên, một kỹ sư lại dễ dàng tránh được những điểm này khi sử dụng các thuật toán tiến hóa để phát triển phần mềm thông minh.

Có thể việc loại bỏ được những hoạt động kém hiệu năng như vừa mô tả sẽ làm giảm được nhiều cấp khuếch đại cho con số 1031 – 1044 FLOPS được tính ở trên, nhưng không may, số cấp khuếch đại cụ thể là rất khó tìm hiểu. Thậm chí, ta cũng khó có thể ước lượng được sơ bộ – hay không biết được các khoản tiết kiệm hiệu năng có thể là năm, 10 hay 25 cấp khuếch đại.¹⁵

Các kiểu tiến hóa như vậy còn gặp phải một tình huống phức tạp khác, trong đó chúng ta khó có thể tìm được từ chúng thậm chí là một cạm bẫy mơ hồ về sự khó khăn của việc tiến hóa trí tuệ. Chúng ta phải tránh suy luận sai lầm rằng vì sự sống thông minh đã tiến hóa trên mặt đất nên các quá trình tiến hóa liên quan có một xác suất ưu

tiên tương đối cao cho việc tạo ra trí thông minh. Suy luận này không chính xác, bởi lẽ nó không tính đến hiệu ứng chọn lọc quan sát.



Hình 3: Hiệu năng của siêu máy tính. Theo nghĩa hẹp, “định luật Moore” quan sát được rằng số bóng bán dẫn trên các bảng mạch tích hợp trong vài thập kỷ qua tăng gấp đôi khoảng hai năm một lần. Tuy nhiên, thuật ngữ này thường thường được sử dụng để nói về một quan sát có tính tổng quát hơn, rằng nhiều tiêu chí hiệu năng trong công nghệ tính toán cũng tuân theo một xu hướng cấp số nhân có tốc độ cao tương tự. Ở đây chúng tôi vẽ đồ thị tốc độ đỉnh của máy tính nhanh nhất thế giới như một hàm số thời gian (theo thang logarit trên trục tung). Trong những năm gần đây, sự gia tăng về tốc độ tuần tự (serial speed) của các bộ xử lý có chậm lại, nhưng việc chuyển đổi song song (parallelization) ngày càng nhiều đã cho phép

*tổng số phép toán thực hiện được tiếp tục bắt kịp với xu hướng trong
đồ thị.¹⁶*

Hiệu ứng này đảm bảo việc toàn bộ người quan sát đều thấy rằng mình được hình thành trên một hành tinh nơi sự sống thông minh được sinh ra, bất kể xác suất của việc một hành tinh như vậy có khả năng tạo ra trí tuệ là bao nhiêu. Chẳng hạn, giả sử rằng ngoài các hiệu ứng có hệ thống của chọn lọc tự nhiên, sự sống thông minh cũng cần có cực kỳ nhiều sự trùng hợp may mắn để ra đời – đủ để sự sống thông minh chỉ tiến hóa trên một trong số 1030 hành tinh mà trên đó, các yếu tố sao chép đơn giản xuất hiện. Trong trường hợp này, khi chạy những thuật toán di truyền để cố gắng sao chép những gì mà tiến hóa tự nhiên đã làm, chúng ta có thể thấy rằng mình phải chạy khoảng 1030 bản mô phỏng trước khi tìm thấy một nơi mà mọi yếu tố được gom đúng cách. Điều này có vẻ hoàn toàn nhất quán với quan sát của chúng tôi rằng sự sống đã tiến hóa trên chính Trái đất. Chỉ thông qua suy luận thận trọng và tương đối phức tạp – bằng cách phân tích những trường hợp hội tụ tiến hóa (convergent evolution) điển hình của các tính trạng có liên quan đến trí tuệ và cố gắng tìm hiểu sự tinh tế của lý thuyết chọn lọc quan sát – thì chúng ta mới có thể phần nào tránh được rào cản nhận thức này. Nếu không chấp nhận thách thức để làm việc đó, ta sẽ không thể loại trừ khả năng “cận trên” giả định về những yêu cầu tính toán nhằm tóm lược quá trình tiến hóa của trí tuệ được nói đến trong Khung 3 có thể sẽ quá thấp, đến 30 cấp khuếch đại (hay một con số nào đó lớn tương tự).¹⁷

Chúng ta cũng có thể lập luận về tính khả thi của trí tuệ nhân tạo bằng cách viện dẫn não người và khuyến nghị rằng mình có thể dùng

nó làm mẫu cho trí tuệ máy. Ta có thể phân biệt những phiên bản khác nhau của cách tiếp cận này dựa trên mức độ tiệm cận của việc bắt chước các chức năng não bộ sinh học mà chúng đề xuất. Ở một thái cực – bắt chước rất giống – chúng ta sẽ có ý tưởng giả lập hoàn chỉnh não bộ được trao đổi trong phần sau. Ở thái cực khác là những cách tiếp cận lấy cảm hứng từ hoạt động chức năng của não bộ nhưng không cố bắt chước ở cấp độ thấp. Những tiến bộ trong khoa học thần kinh và tâm lý nhận thức – vốn sẽ được trợ giúp bởi những cải tiến về công cụ – cuối cùng sẽ khám phá ra những nguyên tắc chung về chức năng não bộ. Những kiến thức này sau đó sẽ trở thành kim chỉ nam cho những nỗ lực về AI. Chúng ta đã có được mạng lưới thần kinh như một ví dụ của kỹ thuật AI lấy cảm hứng từ não bộ. Tổ chức nhận thức phân cấp là một ý tưởng khác được đưa từ ngành khoa học não bộ sang học máy. Nghiên cứu học tăng cường (reinforcement learning) có (ít nhất một phần) động lực từ vai trò của ngành này trong các lý thuyết tâm lý về nhận thức động vật, và các kỹ thuật học tăng cường (ví dụ: “thuật toán TD”) lấy cảm hứng từ những lý thuyết này cũng đang được sử dụng rộng rãi trong AI.¹⁸ Những trường hợp tương tự sẽ được tích lũy nhiều hơn nữa trong tương lai. Vì chỉ có số lượng giới hạn (có thể rất nhỏ) những cơ chế cơ bản khu biệt vận hành trong não bộ, nên toàn bộ chúng cuối cùng sẽ được khám phá nhờ sự tiến bộ liên tục trong ngành khoa học não bộ. Trước khi điều này xảy ra, một cách tiếp cận hỗn hợp, kết hợp một số kỹ thuật lấy cảm hứng từ não bộ và một vài phương pháp nhân tạo thuần túy có thể sẽ cán đích. Trong trường hợp này, hệ thống có được không cần giống não bộ tới mức có thể nhận ra, ngay

cả khi nó được phát triển nhờ sử dụng một số kiến thức thu được từ não bộ.

Sự sẵn có của não bộ để làm mẫu tạo ra một nền tảng hỗ trợ vững chắc cho tuyên bố rằng trí tuệ máy cuối cùng cũng khả thi. Tuy nhiên điều này không cho phép dự báo thời điểm đạt được điều đó, vì tốc độ khám phá trong khoa học não bộ ở tương lai là rất khó đoán. Điều chúng ta có thể nói là: càng nhìn sâu vào tương lai thì khả năng những bí mật của não bộ được giải mã đủ để cho phép tạo ra trí tuệ máy theo cách này càng cao.

Những người khác nhau cùng làm việc để hướng tới trí tuệ máy có những quan điểm khác nhau về mức độ hứa hẹn của những cách tiếp cận hình thái thần kinh (neuromorphic) so với những cách tiếp cận nhắm đến các thiết kế hoàn toàn tổng hợp. Sự tồn tại của những con chim cho thấy các vật nặng hơn không khí có thể thực sự bay được và thúc đẩy nỗ lực chế tạo máy bay, nhưng những chiếc máy bay đầu tiên bay được không hề vỗ cánh. Chúng ta vẫn chưa biết liệu trí tuệ máy có giống như khả năng bay, thứ mà con người đạt được nhờ một cơ chế nhân tạo, hay như sự cháy (mà ban đầu chúng ta làm chủ được nhờ mô phỏng các đám cháy tự nhiên) hay không.

Ý tưởng của Turing về việc thiết kế một chương trình với nội dung kiến thức thụ đắc thông qua học tập thay vì lập trình trước có thể áp dụng một cách tương đương vào các cách tiếp cận trí tuệ máy hình thái học thần kinh và tổng hợp.

Một dị bản của khái niệm máy trẻ con của Turing là ý tưởng về “AI hạt giống” (seed AI).¹⁹ Trong khi cỗ máy trẻ con, như Turing đã hình dung ra, có kiến trúc tương đối cố định và phát triển tiềm năng nội tại

đơn giản bằng cách tích lũy nội dung, một AI hạt giống có thể là một trí tuệ nhân tạo phức tạp hơn với khả năng cải thiện kiến trúc riêng của mình. Trong giai đoạn ban đầu của một AI hạt giống, những cải thiện này xảy ra chủ yếu nhờ các quá trình thử và sai, thụ đắc thông tin, hoặc trợ giúp từ các lập trình viên khác. Tuy nhiên trong những giai đoạn sau, một AI hạt giống có thể đủ hiểu hoạt động riêng của nó để thiết kế những thuật toán và cấu trúc tính toán mới nhằm tự phát triển hiệu năng nhận thức của mình. Sự hiểu biết cần thiết này có thể có được từ việc AI hạt giống đạt đến một cấp độ trí tuệ tổng quát phù hợp trên nhiều lĩnh vực, hoặc từ việc vượt qua một ngưỡng nào đó trong một lĩnh vực liên quan cụ thể (như khoa học máy tính hay toán học).

Điều này đưa chúng ta đến với một khái niệm quan trọng khác, khái niệm “tự cải thiện đệ quy” (recursive self-improvement). Một AI hạt giống thành công sẽ có khả năng lặp lại việc tự cải thiện bản thân: một phiên bản AI ban đầu có thể thiết kế phiên bản cải thiện của chính nó, và phiên bản được cải thiện – đã thông minh hơn bản gốc – có thể thiết kế một phiên bản khác thông minh hơn, và cứ thế.²⁰ Dưới một số điều kiện, quá trình tự cải thiện đệ quy này có thể tiếp diễn đủ lâu để tạo ra một sự bùng nổ trí tuệ, một sự kiện mà trong đó, cấp độ trí tuệ của một hệ thống tăng từ sở hữu các năng lực nhận thức ở mức độ trung bình (có thể dưới cấp độ con người trong hầu hết các khía cạnh, nhưng có tài năng chuyên biệt về lập trình và nghiên cứu AI) lên tới siêu trí tuệ toàn diện chỉ trong thời gian ngắn. Chúng ta sẽ trở lại với khả năng quan trọng này trong Chương 4, trong đó những động lực của một sự kiện như thế sẽ được phân tích cụ thể hơn. Lưu ý rằng mô hình này gợi ý về khả

năng xảy ra những bất ngờ: những cố gắng chế tạo trí tuệ nhân tạo tổng thể có thể sẽ thất bại khá thảm hại cho tới khi cấu phần trọng yếu cuối cùng được đặt đúng chỗ, và đó là thời điểm một AI hạt giống có được khả năng tự cải thiện đệ quy liên tục.

Trước khi kết thúc phần này, còn một điểm nữa chúng ta cần nhấn mạnh: trí tuệ nhân tạo không cần quá giống với trí não con người. AI có thể – thực ra ra rất có thể – vô cùng xa lạ. Chúng ta nên kỳ vọng rằng chúng sẽ có kiến trúc nhận thức khác với trí tuệ sinh học, và trong những giai đoạn phát triển đầu tiên, chúng cũng sẽ có “hồ sơ năng lực” khác biệt về điểm mạnh và điểm yếu trong nhận thức (mặc dù, như chúng tôi sẽ lập luận sau, cuối cùng chúng sẽ có thể vượt qua toàn bộ yếu điểm ban đầu). Hơn nữa, những hệ thống mục tiêu của AI có thể hoàn toàn khác với những hệ thống của loài người. Không có lý do để trông đợi một AI đặc trưng bị kích thích bởi tình yêu, lòng căm thù hay sự tự hào, hay những cảm xúc tương tự của loài người: những thích ứng phức tạp này đòi hỏi nỗ lực quá lớn để tái tạo trong các hệ thống AI. Điều này là một thách thức, đồng thời cũng là một cơ hội lớn. Chúng ta sẽ trở lại vấn đề tạo động lực cho AI trong những chương sau, nhưng điều đó quan trọng với luận điểm trong cuốn sách này đến mức cần luôn được ghi nhớ.

Giả lập hoàn chỉnh não bộ

Trong giả lập hoàn chỉnh não bộ (còn được gọi là “uploading” – tải lên), phần mềm thông minh có thể được tạo ra bằng cách quét và lập mô hình chính xác cấu trúc tính toán của một bộ não sinh học. Vì vậy, một cách nào đó, phương pháp sao chép “trắng trợn” này cũng đại diện cho việc lấy cảm hứng từ tự nhiên. Để giả lập hoàn chỉnh não bộ, ta cần thực hiện những bước đi sau đây.

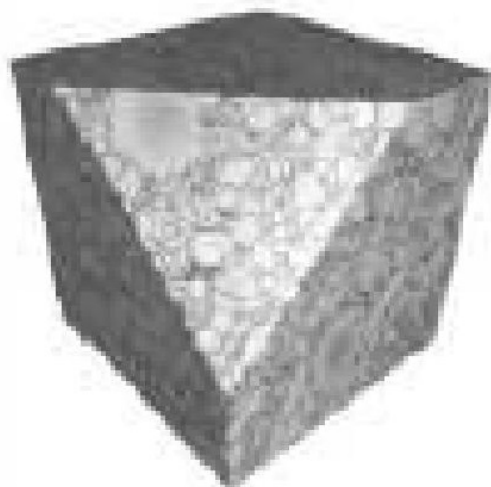
Trước tiên, cần tạo ra một bản quét đầy đủ chi tiết của một bộ não người cụ thể. Việc này có thể liên quan đến thao tác ổn định não sau khi chết nhờ thủy tinh hóa (quá trình biến mô tế bào thành một loại thủy tinh). Một cỗ máy sau đó sẽ cắt mô thành những lát mỏng đưa vừa vào một cỗ máy khác để quét, có thể bởi một hệ thống kính hiển vi điện tử. Những loại thuốc màu khác nhau có thể được sử dụng trong giai đoạn này để làm nổi lên các thuộc tính về cấu trúc và hóa học. Nhiều máy quét có thể cùng làm việc song song để xử lý đồng thời nhiều lát cắt não.

Thứ hai, dữ liệu thô từ các máy quét được đưa vào một máy tính để tự động xử lý hình ảnh nhằm tái tạo mạng lưới thần kinh ba chiều tạo nên nhận thức trong bộ não gốc. Trên thực tế, bước này có thể tiến hành song song với bước đầu tiên để giảm lượng dữ liệu hình ảnh độ phân giải cao cần lưu trong bộ nhớ đệm. Bản đồ thu được sau đó được kết hợp với một thư viện các mô hình điện toán thần kinh của các loại neuron khác nhau hoặc các yếu tố thần kinh khác nhau (ví dụ như một loại kết nối synapse cụ thể). Hình 4 biểu thị một số kết quả quét và xử lý hình ảnh được tạo ra bằng công nghệ ngày nay.

Trong giai đoạn thứ ba, cấu trúc điện toán thần kinh thu được từ bước trước được thực thi trên một máy tính đủ mạnh. Nếu hoàn toàn thành công, ta sẽ thu được một bản sao kỹ thuật số của trí tuệ gốc, với ký ức và nhân cách được giữ nguyên. Trí não giả lập giờ đây tồn tại như một phần mềm máy tính. Trí não này có thể sống trong một thực tại ảo hoặc giao tiếp với thế giới bên ngoài bằng những robot gắn thêm.

Phương pháp giả lập hoàn chỉnh não bộ không đòi hỏi chúng ta phải tìm ra cách hoạt động của nhận thức con người hay cách thức lập trình một trí tuệ nhân tạo, mà chỉ đòi hỏi chúng ta hiểu được các đặc tính chức năng cấp thấp của những phần tử điện toán cơ bản trong não bộ. Để giả lập hoàn chỉnh não bộ thành công, ta không cần bất kỳ đột phá cơ bản nào về khái niệm hoặc lý thuyết.

Tuy nhiên, giả lập hoàn chỉnh não bộ cần có một số công nghệ hỗ trợ khá tiên tiến. Có ba công nghệ tiên quyết quan trọng: (1) quét (scanning): kính hiển vi công suất cao với đủ khả năng phân giải và phát hiện những thuộc tính liên quan; (2) dịch (translation): phân tích hình ảnh tự động để biến dữ liệu quét thô thành một mô hình ba chiều được diễn dịch của các phần tử điện toán thần kinh liên quan; và (3) mô phỏng (simulation): phần cứng đủ mạnh để triển khai cấu trúc điện toán thu được (xem Bảng 4).



Hình 4: Phục dựng giải phẫu thần kinh 3D từ hình ảnh kính hiển vi điện tử. Phía trên bên trái: một vi đồ điện tử tiêu biểu thể hiện mặt cắt của vật chất thần kinh – các sợi nhánh và sợi trục. Phía trên bên phải: Hình ảnh nổi mô thần kinh vống mạc của thỏ lấy được bằng kính hiển vi điện tử quét mặt khối nối tiếp.²¹ Những hình ảnh 2D riêng lẻ được xếp chồng lên nhau thành một khối lập phương (có cạnh khoảng 11 μm). Bên dưới: phục dựng một nhóm phụ neuron lấp đầy thể tích một vùng kết thần kinh (neuropil), được một thuật toán phân đoạn tự động tạo ra.²²

(Khi so sánh với những bước đi khó khăn hơn này, việc xây dựng một thực tại ảo cơ bản hay một cơ thể robot với một kênh nghe nhìn đầu vào và một vài kênh đầu ra là tương đối dễ dàng. Cơ cấu I/O (Đầu vào/đầu ra) đơn giản nhưng phù hợp có vẻ như đã sẵn sàng với công nghệ hiện tại.²³)

Chúng ta có lý do hợp lý để nghĩ rằng những công nghệ hỗ trợ tiên quyết là khả thi, cho dù không phải trong tương lai gần. Những mô hình tính toán hợp lý của nhiều loại neuron và các quá trình thần kinh cũng đã tồn tại. Phần mềm nhận dạng hình ảnh đã được phát triển để có thể đồ lại được các sợi nhánh và sợi trục qua một cụm những hình ảnh hai chiều (mặc dù độ tin cậy vẫn cần được cải thiện). Cũng có những công cụ hình ảnh cung cấp độ phân giải cần thiết – với một kính hiển vi quét xuyên hầm, ta có thể “nhìn thấy” những nguyên tử riêng lẻ, với độ phân giải cao hơn nhiều so với nhu cầu. Tuy nhiên, mặc dù kiến thức và năng lực hiện có cho thấy về mặt nguyên tắc, không có rào cản nào đối với việc phát triển những công nghệ hỗ trợ tiên quyết, nhưng rõ ràng chúng ta vẫn cần thêm rất

nhiều tiến bộ kỹ thuật để giả lập hoàn chỉnh não bộ con người nằm trong tầm với.²⁴ Ví dụ, công nghệ hiển vi sẽ không chỉ cần đến độ phân giải mà còn cả công suất phù hợp. Sử dụng kính hiển vi quét xuyên hầm có độ phân giải cấp nguyên tử để chụp ảnh diện tích mặt phẳng cần thiết sẽ là quá chậm để ứng dụng thực tế. Một kính hiển vi điện tử có độ phân giải thấp hơn sẽ khả dụng hơn, nhưng sẽ cần đến những phương pháp mới để chuẩn bị và nhuộm màu mô vỏ não nhằm hiển thị những chi tiết liên quan, như các cấu trúc synapse tinh vi. Ngoài ra cũng cần một sự mở rộng đáng kể về lượng tư liệu điện toán thần kinh và những cải tiến lớn về xử lý hình ảnh tự động và diễn dịch bản quét (scan interpretation).

Bảng 4 Những năng lực cần thiết để giả lập hoàn chỉnh não bộ

Quét	Xử lý sơ bộ/cố định		Chuẩn bị các bộ não một cách thích hợp, bảo tồn vi cấu trúc và trạng thái liên quan
	Xử lý vật lý		Các phương pháp thao tác trên các bộ não và các mảng mô đã cố định trước, trong và sau khi quét
	Chụp ảnh	Thể tích	Quét toàn bộ thể tích não với thời gian và chi phí hợp lý
		Độ phân giải	Quét với độ phân giải thích hợp để có thể tái tạo

Mô phỏng cơ thể

Mô phỏng cơ thể cho phép tương tác với môi trường ảo hoặc với môi trường thật qua robot

Mô phỏng môi trường

Môi trường ảo cho cơ thể ảo

Nhìn chung, so với trí tuệ nhân tạo, giả lập hoàn chỉnh não bộ ít dựa vào lý thuyết hơn nhưng lại dựa vào năng lực công nghệ nhiều hơn. Mức độ công nghệ cần cho giả lập hoàn chỉnh não bộ phụ thuộc vào mức độ trừu tượng mà não bộ được mô phỏng. Ở khía cạnh này có một sự đánh đổi giữa lý thuyết và công nghệ. Nói chung, khi thiết bị quét và máy tính càng yếu, thì chúng ta càng ít nhờ cậy được vào việc mô phỏng các quá trình hóa học và điện sinh học của não bộ, và càng cần nhiều hiểu biết lý thuyết về kiến trúc tính toán mà mình đang tìm cách mô phỏng để tạo ra những phép biểu diễn trừu tượng hơn của các chức năng liên quan. Ngược lại, với công nghệ quét đủ tiên tiến và công suất tính toán dư thừa, chúng ta có thể xử lý tổng lực một bản giả lập ngay cả khi chỉ có hiểu biết tương đối hạn chế về não bộ. Trong một trường hợp giả tưởng hạn chế, chúng ta có thể hình dung việc giả lập não bộ ở cấp độ các hạt cơ bản, sử dụng phương trình cơ học lượng tử Schrödinger. Khi đó, ta hoàn toàn có thể chỉ dựa vào kiến thức hiện có về vật lý mà không cần đến bất kỳ một mô hình sinh học nào. Tuy nhiên, trường hợp này đặt ra những đòi hỏi vô cùng phi thực tế về công suất tính toán và quá trình thu thập dữ liệu. Mức độ giả lập khả dĩ hơn nhiều kết hợp những neuron riêng lẻ với ma trận kết nối của chúng, cùng với một phần cấu trúc

cây của các sợi nhánh và có thể một vài biến trạng thái của các synapse riêng lẻ. Những phân tử chất dẫn truyền thần kinh sẽ không được mô phỏng riêng lẻ, mà nồng độ không ổn định của chúng sẽ được lập mô hình dạng thô (coarse-grained).

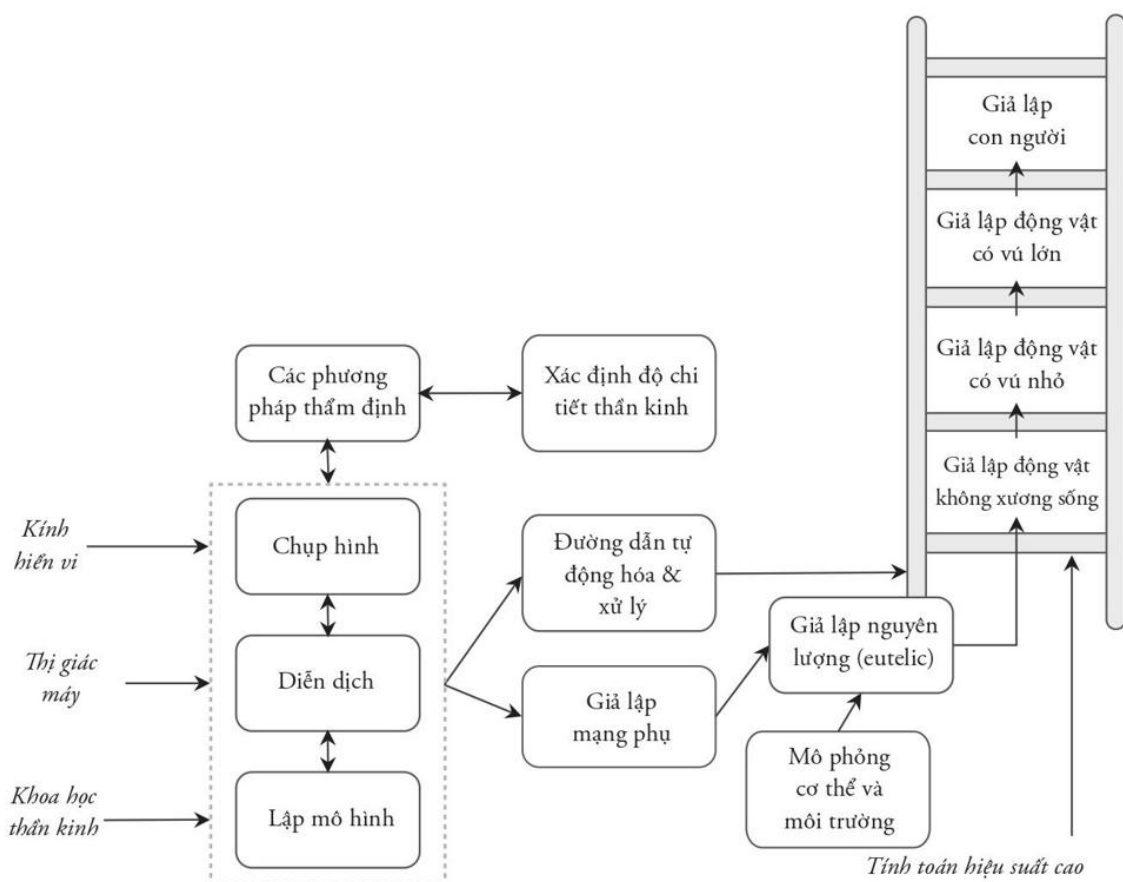
Để đánh giá tính khả thi của việc giả lập hoàn chỉnh não bộ, ta cần hiểu các tiêu chí đánh giá thành công. Mục tiêu không phải là tạo ra một bản mô phỏng não bộ chi tiết và chính xác tới mức sử dụng được để dự báo chính xác những gì có thể xảy ra trong bộ não gốc nếu đặt nó dưới tác động của một chuỗi kích thích. Thay vào đó, mục tiêu là thu thập đầy đủ các đặc tính chức năng về mặt điện toán của não bộ để cho phép kết quả mô phỏng thực hiện được hoạt động trí não. Với mục tiêu này, đa phần các chi tiết sinh học rối rắm của một bộ não thật đều không cần đến.

Một phép phân tích kỹ lưỡng hơn sẽ phân biệt giữa các cấp độ thành công khác nhau của quá trình giả lập, dựa trên mức độ bảo tồn chức năng xử lý thông tin của bộ não được giả lập. Ví dụ, ta có thể phân biệt giữa (1) giả lập độ trung thực cao là tập hợp đầy đủ kiến thức, kỹ năng, năng lực và các giá trị của bộ não được giả lập; (2) giả lập biến dạng, với kết cấu khác biệt đáng kể so với não người nhưng cơ bản vẫn có thể thực hiện hoạt động trí tuệ với tư cách là một bộ não giả lập; (3) mô phỏng chung (cũng có thể bị biến dạng) giống như một đứa trẻ sơ sinh, thiếu những kỹ năng hay ký ức thu được bởi bộ não trưởng thành được giả lập nhưng có khả năng học được hầu hết những gì một người bình thường có thể học.²⁶

Trong khi việc tạo ra bộ não giả lập độ trung thực cao xem ra rất khả thi, rất có thể bản giả lập hoàn chỉnh não bộ đầu tiên mà chúng ta có thể làm được nếu đi theo con đường này sẽ ở một cấp độ thấp

hơn. Mọi thứ sẽ chưa thực sự như ý cho tới khi chúng hoạt động một cách hoàn hảo. Một sự thúc đẩy theo hướng công nghệ giả lập có thể dẫn tới việc tạo ra một kiểu AI hình thái học thần kinh nào đó có ứng dụng một số nguyên tắc tính toán neuron được khám phá trong khi thực hiện giả lập và lai ghép chúng với các phương pháp tổng hợp, và như vậy nó có thể xảy ra trước khi hoàn thành một bản giả lập hoàn chỉnh não bộ đầy đủ chức năng. Xác suất của một hiệu ứng lan tỏa sang AI hình thái học thần kinh như vậy (như chúng ta sẽ xem xét trong một chương dưới đây) làm phức tạp thêm đánh giá chiến lược về mức độ mong muốn của công cuộc theo đuổi khám phá công nghệ mô phỏng.

Còn bao xa chúng ta mới giả lập được hoàn chỉnh não bộ con người? Một đánh giá vừa được thực hiện đã trình bày một lộ trình kỹ thuật và kết luận rằng những năng lực tiên quyết có thể sẽ có sẵn vào khoảng giữa thế kỷ này, tuy với một sai số bất định lớn.²⁷ Hình 5 biểu thị những cột mốc chính trong lộ trình này.



Hình 5 Lộ trình giả lập hoàn chỉnh não bộ. Sơ đồ đầu vào, hoạt động và cột mốc.²⁸

Về ngoài đơn giản của sơ đồ có thể đánh lừa con mắt, tuy nhiên chúng ta cần thận trọng và không nên đánh giá quá thấp lượng công việc còn cần phải làm. Hiện ta chưa giả lập được bộ não nào cả. Hãy xem xét một sinh vật mô hình khiêm tốn có tên khoa học là *Caenorhabditis elegans*, loài giun tròn trong suốt, chỉ dài 1 mm và có 302 neuron. Ma trận kết nối hoàn chỉnh của những neuron này đã được biết đến từ giữa những năm 1980, khi nó được vất vả vẽ ra bằng mọi biện pháp như tách lớp, hiển vi điện tử và gán nhãn thủ

công cho các mẫu vật.²⁹ Nhưng việc chỉ biết neuron nào kết nối với neuron nào vẫn là không đủ. Để tái tạo mô phỏng não bộ ta cũng cần biết các synapse nào là phản xạ, các synapse nào là ức chế; cường độ của các kết nối và những thuộc tính động khác nhau của các sợi trục, synapse và cây sợi nhánh. Thông tin này vẫn chưa sẵn có, thậm chí đối với hệ thần kinh nhỏ của *C. elegans* (tuy hiện tại nó có thể nằm trong phạm vi của một dự án nghiên cứu mục tiêu có quy mô vừa phải).³⁰ Thành công trong việc giả lập một bộ não nhỏ xíu như của *C. elegans* có thể cho chúng ta biết rõ hơn những việc làm để giả lập những bộ não lớn.

Tại một thời điểm nào đó trong quá trình phát triển công nghệ, một khi các kỹ thuật đã sẵn có để tự động giả lập những lượng mô não nhỏ, ta sẽ chỉ còn lại vấn đề về quy mô. Hãy lưu ý “cái thang” bên phải trong Hình 5. Chuỗi các ô lên cao dần đại diện cho một chuỗi chung cuộc các bước tiến có thể thực hiện sau khi những rào cản đầu tiên đã dỡ bỏ. Các giai đoạn trong chuỗi này tương ứng với các bản giả lập hoàn chỉnh não bộ của các sinh vật mô hình nối tiếp nhau với độ phức tạp về mặt thần kinh tăng dần – ví dụ: *C. elegans* → ong mật → chuột → khỉ rhesus → con người. Vì khoảng cách giữa những bậc thang này – ít nhất là sau bậc đầu tiên – hầu hết có bản chất định lượng và chủ yếu (mặc dù không hoàn toàn) là do sự khác biệt về kích thước não cần mô phỏng, nên chúng sẽ có thể được xử lý dễ dàng bằng việc mở rộng quy mô một cách tương đối đơn giản của khả năng quét và mô phỏng.³¹

Khi chúng ta bắt đầu leo lên những bậc thang này, việc có được bản giả lập hoàn chỉnh não bộ của con người ở giai đoạn cuối rõ ràng là có thể thấy trước được.³² Do đó chúng ta có thể kỳ vọng về

việc có được một vài cảnh báo sớm trước khi với tới trí tuệ máy cấp độ con người, ít nhất nếu công nghệ quét công suất cao hay công suất tính toán cần cho mô phỏng thời gian thực là công nghệ cuối cùng trong số những công nghệ tiên quyết đạt đến mức độ phát triển thích hợp. Tuy nhiên, nếu công nghệ hỗ trợ cuối cùng có được lại là lập mô hình điện toán thần kinh thì việc chuyển đổi từ những nguyên mẫu không ấn tượng thành bản giả lập con người vận hành được có thể sẽ gian nan hơn. Ta có thể hình dung ra một kịch bản, trong đó, bất chấp việc dữ liệu quét và máy tính tốc độ cao là thừa thãi, thì mô hình thần kinh vận hành đúng vẫn là một thứ rất khó đạt được. Khi rào cản cuối cùng đã được loại bỏ, thứ mà trước kia là một hệ thống đã hỏng hoàn toàn – bạn có thể nghĩ đến một bộ não vô thức đang trải qua một cơn tai biến nặng – có thể biến thành một trạng thái thức tỉnh nhất quán. Trong trường hợp này, bước tiến then chốt sẽ không được báo trước bởi một loạt các bản giả lập động vật hoạt động tốt với độ lớn tăng dần (như những dòng tí giọt gân trên báo với kích thước phông chữ lớn dần). Thậm chí đối với những người quan tâm, cũng khó có thể nói trước còn bao nhiêu thiếu sót trong những mô hình điện toán thần kinh tại từng thời điểm và phải mất bao nhiêu lâu để xử lý chúng, thậm chí ngay vào đêm trước khi xảy ra những đột phá quan trọng. (Một khi việc giả lập hoàn chỉnh não người đã được thực hiện, những sự phát triển có tiềm năng bùng nổ kế tiếp có thể sẽ xảy ra; nhưng chúng ta sẽ bàn sau trong Chương 4.)

Vì thế, chúng ta hoàn toàn có thể hình dung ra được những kịch bản vốn bất ngờ của việc giả lập hoàn chỉnh não bộ, thậm chí nếu toàn bộ các nghiên cứu liên quan được thực hiện công khai. Mặc dù vậy, so với con đường đến với trí tuệ máy của AI, việc giả lập hoàn

chính não bộ sẽ có “điềm báo” rõ ràng hơn vì nó trông cậy nhiều hơn vào những công nghệ cụ thể có thể quan sát được và không hoàn toàn dựa trên những hiểu biết lý thuyết. Chúng ta cũng có thể nói, với mức độ tự tin cao hơn so với con đường AI, rằng việc giả lập này sẽ không thành công trong tương lai gần (chẳng hạn trong 50 năm tới), bởi lẽ chúng ta biết rằng một số công nghệ tiền thân đầy thách thức vẫn chưa được phát triển. Ngược lại, một ai đó, về nguyên tắc, xem ra có thể ngồi xuống và lập trình ra một AI hạt giống trên một chiếc máy tính bình dân hiện có; và ta có thể tưởng tượng được – cho dù không mấy chắc chắn – rằng một người ở đâu đó sẽ có một góc nhìn đúng đắn về phương pháp thực hiện điều này trong tương lai gần.

Nhận thức sinh học

Con đường thứ ba dẫn đến trí tuệ cao hơn con người hiện tại là tăng cường chức năng của các bộ não sinh học. Về nguyên tắc, điều này có thể đạt được thông qua sinh sản chọn lọc, mà không cần đến công nghệ. Tuy nhiên, bất cứ cố gắng khởi tạo một chương trình ưu sinh quy mô lớn cổ điển nào cũng đều phải đối mặt với những rào cản chính trị và đạo đức. Hơn nữa, trừ khi việc chọn lọc là cực mạnh, nếu không ta sẽ cần đến nhiều thế hệ để tạo ra những kết quả đáng kể. Trước khi một dự án như vậy đơm hoa kết trái, những tiến bộ trong công nghệ sinh học có thể cho phép kiểm soát các đặc điểm di truyền và sinh học thần kinh của con người theo cách trực tiếp hơn nhiều, khiến bất cứ chương trình sinh sản nào của con người cũng trở nên vô ích. Vì vậy, chúng tôi tập trung vào những phương pháp có tiềm năng đem lại kết quả nhanh hơn, với thang thời gian chỉ trong một vài thế hệ hoặc ít hơn.

Những năng lực nhận thức riêng lẻ của chúng ta có thể được củng cố bằng nhiều cách khác nhau, bao gồm cả những phương pháp truyền thống như giáo dục đào tạo. Sự phát triển về thần kinh có thể được thúc đẩy bởi những biện pháp can thiệp “low-tech” (không cần nhiều công nghệ) như tối ưu hóa dinh dưỡng bà mẹ trẻ em, loại bỏ chì và những chất ô nhiễm độc hại cho thần kinh ra khỏi môi trường, diệt ký sinh trùng, bảo đảm chế độ ngủ và luyện tập thích hợp và phòng chống những bệnh dịch ảnh hưởng đến não bộ.³³ Những cải thiện về nhận thức có thể đạt được nhờ từng biện pháp nói trên, mặc dù biên độ tăng có thể chỉ ở mức khiêm tốn, đặc biệt đối với những người đã được nuôi dưỡng và giáo dục tương đối tốt. Chắc chắn chúng ta sẽ không đạt được siêu trí tuệ bằng bất cứ biện pháp nào trong số đó, nhưng chúng vẫn có đôi chút hữu ích, cụ thể là vực dậy những người yếu thế và mở rộng nhân tài toàn cầu. (Suy giảm trí tuệ suốt đời do thiếu i-ốt vẫn còn phổ biến tại nhiều khu vực nghèo khó trên thế giới – đây là điều đáng phẫn nộ vì chứng bệnh này có thể được phòng ngừa bằng muối ăn bổ sung i-ốt với chỉ một vài cent một năm một người.³⁴)

Những cải tiến y sinh có thể mang lại những cú hích mạnh mẽ hơn. Đã có những loại thuốc được coi là tăng cường trí nhớ, sự tập trung và năng lượng thần kinh, ít nhất là cho một vài đối tượng.³⁵ (Tôi viết cuốn sách này nhờ nguồn “nhiên liệu” từ cà phê và kẹo cao su chứa nicotin). Trong khi tính hiệu quả của thế hệ thuốc thông minh hiện nay là không ổn định, không đáng kể và nhìn chung mơ hồ, các loại thuốc thông minh tương lai lại có thể mang đến những lợi ích rõ ràng hơn với ít hiệu ứng phụ hơn.³⁶ Tuy nhiên, trên cả hai nền tảng thần kinh học và tiến hóa, ít có khả năng là ai đó có thể chiêm ngòi

cho một sự trỗi dậy mạnh mẽ của trí tuệ bằng cách đưa một hóa chất nào đó vào não bộ của một người khỏe mạnh.³⁷ Chức năng nhận thức của não người phụ thuộc vào sự kết hợp tinh vi của nhiều yếu tố, đặc biệt là trong những giai đoạn thiết yếu của quá trình phát triển phôi thai – và chúng ta cần phải cân bằng, tinh chỉnh và nuôi dưỡng cẩn thận để có thể tăng cường sức mạnh của cấu trúc tự tổ chức này, thay vì chỉ ngâm nó vào trong một thứ thuốc ngoại lai nào đó.

Vận dụng các đặc điểm di truyền sẽ cung cấp một bộ công cụ mạnh hơn nhiều so với dược học tâm thần. Hãy xem xét một lần nữa ý tưởng chọn lọc di truyền: thay vì cố triển khai một chương trình ưu sinh bằng cách kiểm soát các mẫu hình giao phối (mating pattern), người ta có thể áp dụng việc chọn lọc ở cấp độ bào thai hay giao tử.³⁸ Chẩn đoán di truyền tiền cấy ghép (pre-implantation genetic diagnosis) đã được áp dụng trong quá trình thụ tinh nhân tạo để sàng lọc những bào thai được tạo ra, nhằm phát hiện những rối loạn đơn gen (monogenic) như bệnh Huntington và khuynh hướng mắc một số căn bệnh khởi phát muộn như ung thư vú. Nó cũng được sử dụng trong chọn lọc giới tính và khớp loại kháng nguyên bạch cầu với các anh chị em bị bệnh, những người sau này có thể hưởng lợi từ việc hiến tế bào gốc máu cuống rốn khi đứa trẻ ra đời.³⁹ Số lượng các tính trạng có thể được chọn hay loại bỏ sẽ gia tăng đáng kể trong một hay hai thập kỷ tới. Một động lực mạnh mẽ của tiến bộ trong di truyền học hành vi là chi phí giảm nhanh chóng của việc xác định kiểu gen và lập trình tự gen. Việc phân tích tính trạng phức tạp trên bộ gen, sử dụng những nghiên cứu với số đối tượng cực lớn, chỉ mới bắt đầu trở nên khả thi và sẽ khiến tri thức của chúng ta về kiến trúc di truyền ở các tính trạng nhận thức và hành vi con người gia

tăng đáng kể.⁴⁰ Bất cứ tính trạng nào với khả năng di truyền tương đối – bao gồm khả năng nhận thức – sẽ trở nên dễ chọn lọc.⁴¹ Chọn lọc bào thai không cần đến hiểu biết sâu sắc về những con đường mang tính nhân quả mà theo đó, các gen tạo ra các kiểu hình trong hoạt động tương tác phức tạp với môi trường: nó chỉ cần (rất nhiều) dữ liệu về các yếu tố tương quan di truyền của các tính trạng được chú ý đến.

Có thể ước lượng sơ bộ về mức độ gia tăng có được từ những kịch bản chọn lọc khác nhau.⁴²

Bảng 5 IQ tối đa gia tăng từ chọn lọc trong một nhóm bào thai⁴³

Chọn lọc	Điểm IQ gia tăng
1 trong 2	4,2
1 trong 10	11,5
1 trong 100	18,8
1 trong 1.000	24,3
5 thế hệ của 1 trong 10	< 65 (do hiệu suất giảm dần)
10 thế hệ của 1 trong 10	< 130 (do hiệu suất giảm dần)
Giới hạn lũy tích (các biến bổ sung được tối ưu hóa cho nhận thức)	100+ (<300 do hiệu suất giảm dần)

Bảng 5 cho thấy những gia tăng kỳ vọng về trí tuệ có được từ những lượng chọn lọc khác nhau, giả định rằng có đầy đủ thông tin về những biến thể di truyền bổ sung hỗ trợ khả năng di truyền theo nghĩa hẹp của trí tuệ. (Với thông tin một phần, hiệu quả của việc chọn lọc sẽ suy giảm, mặc dù không hẳn tới mức độ mà ai đó có thể

ngây thơ kỳ vọng.⁴⁴⁾ Không có gì ngạc nhiên, việc chọn lọc với số lượng phôi thai lớn hơn cho mức điểm gia tăng cao hơn nhưng lại thu được hiệu suất giảm dần mức độ cao: chọn lọc trong số 100 phôi tạo ra một mức gia tăng thấp hơn rất nhiều so với 50 lần của kết quả thu được khi chọn lọc giữa hai phôi.⁴⁵⁾

Điều thú vị là sự suy giảm hiệu quả đã bớt đi đáng kể khi quá trình chọn lọc được trải rộng trên nhiều thế hệ. Do đó, việc lựa chọn cá thể ở vị trí số 1 trên 10 lặp đi lặp lại qua 10 thế hệ (mỗi thế hệ mới bao gồm con cháu của những người được chọn lọc trong thế hệ trước) sẽ tạo ra mức gia tăng lớn hơn nhiều trong giá trị của tính trạng so với việc lựa chọn 1/100 tiến hành một lần duy nhất. Vấn đề của lựa chọn tuần tự, tất nhiên là thời gian. Nếu mỗi thế hệ kéo dài khoảng 20 hay 30 năm, thì thậm chí chỉ năm thế hệ kế tiếp nhau đã đưa chúng ta sang thế kỷ 22, và nhiều phương pháp kỹ thuật mạnh mẽ, trực tiếp hơn (chưa kể đến trí tuệ máy) có lẽ đã xuất hiện trước thời điểm đó rất lâu rồi.

Tuy nhiên, có một công nghệ hỗ trợ, mà khi đã được phát triển để sử dụng trên người, sẽ có thể làm gia tăng đáng kể năng lực cải thiện của quá trình sàng lọc di truyền tiền cấy ghép, cụ thể là tạo ra những tinh trùng và trứng hữu dụng từ tế bào phôi gốc.⁴⁶⁾ Những kỹ thuật để làm việc này đã được sử dụng nhằm tạo ra thế hệ con cháu có khả năng sinh sản ở chuột và các tế bào giống như giao tử ở người. Tuy nhiên, thách thức khoa học lớn vẫn còn tồn tại trong việc chuyển dịch những kết quả trên động vật sang con người và phòng ngừa những bất thường mang tính biểu sinh trong những dòng tế bào gốc được tạo ra. Theo một chuyên gia, những thách thức này có

thể khiến việc ứng dụng trên người chậm lại “10 hoặc thậm chí 50 năm trong tương lai.”⁴⁷

Với những giao tử sinh ra từ tế bào gốc, năng lực chọn lọc mà một cặp đôi có được có thể sẽ gia tăng đáng kể. Trong thực tế hiện tại, một quy trình thụ tinh trong ống nghiệm (IVF) thường tạo ra không quá 10 phôi. Cũng với những giao tử sinh ra từ tế bào gốc, một số ít tế bào được hiến có thể trở thành một lượng giao tử gần như vô hạn mà sau đó có thể kết hợp để tạo ra phôi. Kế tiếp, các phôi này có thể được xác định kiểu hoặc sắp xếp theo trình tự, và những phôi nhiều hứa hẹn nhất sẽ được chọn cấy. Tùy vào chi phí chuẩn bị và sàng lọc từng phôi riêng lẻ, công nghệ này có thể làm gia tăng năng lực chọn lọc cho các cặp đôi sử dụng phương pháp thụ tinh trong ống nghiệm nhiều lần.

Quan trọng hơn, các giao tử sinh ra từ tế bào gốc có thể rút ngắn thời gian chọn lọc kéo dài nhiều thế hệ xuống còn ít hơn thời gian trưởng thành của một người khi cho phép chọn lọc phôi lặp lại (iterated embryo selection). Quy trình này như sau:

1. Xác định kiểu gen và chọn một số lượng phôi có các đặc điểm di truyền mong muốn nổi trội hơn.
2. Chiết xuất tế bào gốc từ những phôi này và biến chúng thành tinh trùng và trứng, trưởng thành trong thời gian dưới sáu tháng.⁴⁹
3. Thụ tinh cho trứng để tạo phôi.
4. Lặp lại cho tới khi những thay đổi di truyền lớn được tích lũy.

Bằng cách này, việc chọn lọc trong 10 thế hệ hoặc nhiều hơn có thể hoàn thành chỉ trong vài năm. (Quy trình này có thể mất nhiều thời gian và đắt đỏ; tuy nhiên, về nguyên tắc, nó có thể được thực hiện chỉ một lần thay vì lặp lại cho mỗi lần sinh. Các dòng tế bào

được thiết lập ở cuối quy trình có thể được dùng để tạo ra những lượng phôi đã được tăng cường rất lớn.)

Như đã thấy trong Bảng 5, mức trung bình của trí tuệ trong số những cá thể được thụ thai theo cách này có thể rất cao, có khả năng ngang với hoặc cao hơn một chút so với cá thể thông minh nhất trong lịch sử loài người. Một thế giới có số lượng lớn cá thể như vậy sẽ có thể (nếu nó có văn hóa, giáo dục, hạ tầng thông tin... tương ứng) tạo thành một siêu trí tuệ tập thể (collective superintelligence).

Tác động của công nghệ này sẽ bị giảm và đình trệ vì một số yếu tố. Có một độ trễ trưởng thành không thể tránh khỏi khi những phôi được lựa chọn cuối cùng trải qua giai đoạn sinh trưởng: cần ít nhất 20 năm trước khi một đứa trẻ được tăng cường như vậy phát triển đầy đủ, và còn lâu hơn nữa cho tới khi những đứa trẻ này tạo thành một phân khúc đáng kể trong lực lượng lao động. Hơn nữa, thậm chí sau khi công nghệ đã được hoàn thiện, tỷ lệ ứng dụng ban đầu có thể chưa cao. Một số quốc gia có thể cấm sử dụng nó hoàn toàn vì lý do đạo đức hay tôn giáo,⁵⁰ và thậm chí ở nơi được phép chọn lọc, nhiều cặp đôi cũng có thể lựa chọn phương thức thụ thai tự nhiên. Tuy nhiên, mong muốn thụ tinh trong ống nghiệm sẽ tăng lên nếu có những lợi ích rõ ràng hơn gắn với phương pháp này – ví dụ như một đảm bảo ngầm rằng đứa trẻ sẽ rất có tài và tránh được khuynh hướng mắc bệnh di truyền. Chi phí chăm sóc sức khỏe thấp hơn và tuổi thọ dự kiến cao hơn cũng có thể là những luận điểm bảo vệ chọn lọc di truyền. Khi việc sử dụng phương pháp này trở nên phổ biến hơn, đặc biệt trong giới thượng lưu, có thể sẽ có một sự dịch chuyển văn hóa hướng tới các tiêu chuẩn làm cha mẹ, trong đó coi việc sử

dụng chọn lọc là điều nên làm đối với các cặp đôi có trách nhiệm và đã được “khai sáng”. Nhiều cặp đôi lúc đầu chưa tha thiết có thể sẽ lại tham gia để có một đứa con không bị kém thế so với những đứa con được tăng cường khác của bạn bè hay đồng nghiệp. Một số quốc gia có thể đưa ra những phần thưởng nhằm khích lệ công dân tận dụng chọn lọc di truyền để gia tăng nguồn vốn nhân lực cho đất nước, hoặc để tăng tính ổn định xã hội dài hạn bằng cách chọn lọc các tính trạng như ngoan ngoãn, vâng lời, tuân thủ, sợ rủi ro hoặc hèn nhát cho những người không thuộc các phe phái cầm quyền.

Những hiệu ứng tác động lên năng lực trí tuệ cũng có thể phụ thuộc vào mức độ mà năng lực chọn lọc sẵn có được dùng để cải thiện các tính trạng nhận thức (Bảng 6). Những người lựa chọn sử dụng một dạng chọn lọc phôi nào đó sẽ phải chọn cách phân bổ năng lực chọn lọc, và trí tuệ ở một mức độ nào đó sẽ cạnh tranh với những thuộc tính mong muốn khác, như thể trạng, sắc đẹp, tính cách hoặc sức mạnh. Chọn lọc phôi lặp lại, nhờ cung cấp một năng lực chọn lọc lớn, có thể giảm bớt gánh nặng trong một số trường hợp đánh đổi, cho phép đồng thời chọn lọc được nhiều tính trạng. Tuy nhiên, quy trình này có xu hướng phá vỡ mối quan hệ di truyền bình thường giữa cha mẹ và con, điều có thể ảnh hưởng xấu đến nhu cầu trong nhiều nền văn hóa.⁵¹

Cùng với những bước tiến xa hơn trong công nghệ di truyền, việc tổng hợp các bộ gen theo chỉ tiêu kỹ thuật có thể trở nên khả thi, loại bỏ đi nhu cầu về những kho chứa phôi lớn. Tổng hợp ADN đã là một công nghệ sinh học phổ thông và tự động hóa phần lớn, mặc dù vẫn chưa thể tổng hợp được một bộ gen người hoàn chỉnh có thể được

sử dụng trong bối cảnh sinh sản (cụ thể là vì những khó khăn chưa giải quyết được trong quá trình tìm hiểu về di truyền học biểu sinh).⁵²

Nhưng khi công nghệ này đã trở nên hoàn thiện, một phôi có thể được thiết kế với tổ hợp mong muốn chính xác của các đầu vào di truyền từ mỗi bên bố hoặc mẹ. Gen của bố hoặc mẹ có thể được cắt nối, bao gồm cả các allele (gen đẳng vị) hiện diện với tần suất thấp trong dân số nhưng có thể có hiệu ứng tích cực đáng kể lên nhận thức.⁵³

Một cách can thiệp nữa sẽ trở nên khả thi khi bộ gen người có thể được tổng hợp, đó là việc “kiểm tra chính tả” (spell-checking) di truyền cho một phôi. (Chọn lọc phôi lặp lại có thể cho phép thực hiện một quá trình gần như vậy). Mỗi người trong chúng ta mang theo một gánh nặng đột biến (mutational load), với khả dĩ hàng trăm đột biến làm giảm hiệu năng của những quá trình tế bào khác nhau.⁵⁴ Mỗi đột biến riêng lẻ có hiệu ứng gần như có thể bỏ qua (do đó nó chỉ bị loại bỏ khỏi vốn gen một cách từ từ), nhưng nếu được tổ hợp, những đột biến này có thể ảnh hưởng nặng nề đến hoạt động của chúng ta.⁵⁵ Những khác biệt riêng lẻ trong trí tuệ rất có thể xuất phát từ những biến thể về lượng và chất của những allele tương đối nguy hại mà mỗi chúng ta đều có. Với việc tổng hợp gen, chúng ta có thể dựa vào bộ gen của một phôi và tạo ra một phiên bản không có sự “nhiều động” di truyền (genetic noise) của các dạng đột biến tích lũy. Nếu muốn phát biểu một cách khiêu khích, một người có thể nói rằng những cá nhân được tạo ra từ những bộ gen đã được “sửa lỗi chính tả” này có thể “ngời” hơn bất cứ người nào đang sống, vì họ có hình dạng con người ít “méo mó” hơn. Những người này không phải là những bản sao chính xác, vì con người khác nhau về mặt di truyền

ở nhiều khía cạnh chứ không chỉ bởi họ mang những đột biến nguy hại khác nhau.

Bảng 6 Những tác động có thể xảy ra từ quá trình chọn lọc di truyền trong các kịch bản khác nhau ⁵⁶				
Tỷ lệ áp dụng/ công nghệ	"IVF+" Chọn lọc một trên hai phôi [4 điểm]	"Aggressive IVF" (IVF tích cực) Chọn lọc một trên 10 phôi [12 điểm]	"In vitro egg" (Trứng trong ống nghiệm) Chọn lọc một trên 100 phôi [19 điểm]	"Iterated embryo selection" (Chọn lọc phôi lặp lại) [100+ điểm]
"Phương pháp sinh sản phụ" ~ 0.25% áp dụng	Không đáng kể về mặt xã hội trong một thế hệ. Ảnh hưởng của bất đồng xã hội quan trọng hơn những tác động trực tiếp.	Không đáng kể về mặt xã hội trong một thế hệ. Ảnh hưởng của bất đồng xã hội quan trọng hơn những tác động trực tiếp.	Số lượng nhỏ các dạng tăng cường ngẫu nhiên có thể nhận ra trong các trường hợp chọn lọc có tính nhận thức cao.	Các cá thể được lựa chọn trở thành các nhà khoa học, luật sư, bác sĩ, kỹ sư ưu tú. Phải chăng là một cuộc "Phục hưng Trí tuệ"?
"Phương pháp xa xỉ" 10% áp dụng	Tác động nhẹ về mặt nhận thức ở thế hệ thứ nhất, kết hợp với việc chọn lọc cho các tính trạng phi nhận thức nhằm mang lại lợi thế đáng kể cho một nhóm thiểu số.	Phần lớn sinh viên đại học ở Harvard đã được tăng cường. Thế hệ thứ hai chiếm ưu thế trong các ngành nghề đòi hỏi nhận thức cao.	Các cá thể được lựa chọn trở thành các nhà khoa học, luật sư, bác sĩ, kỹ sư ưu tú ở thế hệ thứ nhất.	"Nền văn minh hậu nhân loại" (Posthumanity) ⁵⁷
"Phương pháp mới thông thường" >90% áp dụng	Tình trạng mất khả năng học ít phổ biến hơn rất nhiều ở trẻ em. Ở thế hệ thứ hai, số lượng cá thể có IQ cao hơn các ngưỡng trần đã tăng hơn gấp đôi.	Gia tăng đáng kể về trình độ học vấn và thu nhập. Thế hệ thứ hai tăng gấp nhiều lần, phân phối về phía đuôi phải (right tail).	Chỉ số IQ ước tính đặc thù của các nhà khoa học xuất sắc trở nên phổ biến gấp hơn 10 lần ở thế hệ thứ nhất và hàng nghìn lần ở thế hệ thứ hai.	"Nền văn minh hậu nhân loại"

Nhưng biểu thị kiểu hình của một bộ gen đã được sửa lỗi chính tả có thể là một cấu trúc thể chất và tinh thần siêu việt, với hoạt động được cải thiện trong những chiều kích tính trạng đa gen như trí tuệ, sức khỏe, sức chịu đựng và ngoại hình.⁵⁸ (Chúng ta có thể tạm liên tưởng đến những khuôn mặt tổng hợp, trong đó các khiếm khuyết trên từng khuôn mặt riêng lẻ sẽ được cân đối khi xếp chồng lên nhau: xem Hình 6.)



Hình 6: Những khuôn mặt tổng hợp giống như những bộ gen đã “sửa lỗi chính tả”. Mỗi bức hình trung tâm được tạo ra bằng cách xếp chồng những tấm hình của 16 cá nhân khác nhau (người dân Tel Aviv). Những khuôn mặt tổng hợp thường được đánh giá là đẹp hơn so với bất cứ khuôn mặt riêng lẻ nào đã được dùng để tạo ra chúng, vì những điểm đặc tính không hoàn hảo đã được cân đối. Tương tự, bằng cách loại bỏ những đột biến riêng lẻ, các bộ gen được “sửa lỗi chính tả” có thể tạo ra những người gần với “hình thái lý tưởng của Plato”. Những cá nhân này có thể không hề giống nhau về di truyền, vì nhiều gen có nhiều allele với chức năng tương đương. Việc kiểm tra chính tả chỉ có thể loại bỏ những biến dị xuất phát từ các đột biến nguy hại.⁵⁹

Những kỹ thuật công nghệ sinh học tiềm năng khác cũng có thể có liên quan. Nhân bản vô tính người, một khi đã thực hiện được, có thể tái tạo bộ gen của những cá thể đặc biệt xuất chúng. Trí thông minh sẽ bị giới hạn bởi ý muốn có được một sự liên quan về mặt sinh học với con cái của hầu hết các bậc cha mẹ trong tương lai, nhưng phương pháp này vẫn có thể có tác động đáng kể vì (1) ngay cả một sự gia tăng tương đối nhỏ về số lượng người đặc biệt tài giỏi cũng

có ảnh hưởng đáng kể; và (2) có thể một quốc gia nào đó sẽ bắt đầu một chương trình ưu sinh quy mô lớn, có thể bằng cách trả tiền cho các bà mẹ mang thai hộ. Các loại kỹ thuật di truyền khác – chẳng hạn như thiết kế gen tổng hợp mới hoặc chèn vào bộ gen những vùng xúc tác và những phần tử khác để kiểm soát biểu hiện gen – cũng có thể trở nên quan trọng theo thời gian. Thậm chí, những khả năng kỳ lạ cũng có thể tồn tại, ví dụ như những các thùng lớn chứa đầy mô vỏ não được nuôi cấy có cấu trúc phức tạp, hoặc các loài động vật biến đổi gen được “nâng cấp” (chẳng hạn như một số loài động vật có vú não lớn như cá voi hoặc voi được bổ sung gen người). Những loài vừa nhắc tới hoàn toàn là tưởng tượng, nhưng trong thời gian dài, chúng ta cũng không thể hoàn toàn bỏ quên chúng.

Chúng ta đã trao đổi về những liệu pháp can thiệp dòng mầm (germline) có thể được thực hiện trên giao tử hay phôi. Bằng cách bỏ qua chu kỳ thế hệ, các liệu pháp tăng cường gen sinh dưỡng (somatic gene enhancement) về nguyên tắc có thể tạo ra các tác động một cách nhanh chóng hơn, nhưng về mặt công nghệ lại khó thực hiện hơn rất nhiều. Các liệu pháp này đòi hỏi chèn các gen đã được sửa đổi vào một số lượng lớn tế bào trong cơ thể sống – và trong trường hợp tăng cường nhận thức, bao gồm cả não bộ. Việc học lọc từ những tế bào trứng hay phôi sẵn có, ngược lại, không yêu cầu chèn gen. Ngay cả những liệu pháp dòng mầm liên quan đến việc sửa đổi bộ gen (như kiểm tra chính tả bộ gen hay cắt nối các allele hiếm) cũng có thể được triển khai dễ dàng hơn nhiều trong giai đoạn giao tử hay phôi, trong đó ta chỉ cần xử lý một số lượng tế bào nhỏ. Hơn nữa, các liệu pháp can thiệp dòng mầm trên phôi có thể đạt được hiệu quả cao hơn so với can thiệp sinh dưỡng trên người lớn,

vì có thể định hình những phát triển não bộ sớm, trong khi các liệu pháp sinh dưỡng lại có thể chỉ giới hạn trong việc chỉnh đốn một cấu trúc hiện có. (Một phần của những gì có thể được thực hiện nhờ liệu pháp gen sinh dưỡng cũng có thể đạt được nhờ các phương pháp được lý.)

Như vậy, khi tập trung vào các liệu pháp can thiệp dòng mầm, chúng ta phải tính đến độ trễ về thể hệ đang làm chậm bất cứ thay đổi lớn nào trên thế giới.⁶⁰ Thậm chí nếu công nghệ đã hoàn thiện và được sử dụng ngay hôm nay, cũng sẽ phải mất hơn hai thập kỷ để thể hệ được cải thiện gen trưởng thành. Hơn nữa, với những ứng dụng trên người, thường có độ trễ ít nhất một thập kỷ giữa lúc chứng minh ý tưởng trong phòng thí nghiệm và ứng dụng lâm sàng vì cần nghiên cứu kỹ lưỡng để đảm bảo an toàn. Tuy nhiên, những dạng chọn lọc gen đơn giản nhất về cơ bản lại có thể loại bỏ quá trình thử nghiệm này, vì chúng sẽ sử dụng các kỹ thuật xử lý thụ tinh tiêu chuẩn và thông tin di truyền để chọn lọc giữa các phôi mà nếu không làm vậy có thể sẽ được chọn lọc ngẫu nhiên.

Độ trễ cũng có thể phát sinh từ những cản trở không xuất phát từ nỗi sợ thất bại (nhu cầu kiểm nghiệm an toàn) mà từ nỗi sợ thành công – sự cần thiết của việc quản lý được thúc đẩy bởi những mối quan ngại dành cho khả năng được chấp nhận về mặt đạo đức của chọn lọc di truyền hay các hệ lụy xã hội rộng hơn mà phương pháp này mang lại. Những quan ngại này có ảnh hưởng lớn hơn ở một số quốc gia so với những quốc gia khác vì hoàn cảnh văn hóa, lịch sử và tôn giáo khác nhau. Chẳng hạn, nước Đức thời hậu chiến đã chọn cách tránh xa các phương pháp sinh sản nhắm tới việc tăng cường năng lực, kể cả ở mức độ liên quan nhỏ nhất. Đây là một lập trường

dễ hiểu, khi mà quốc gia này đã từng trải qua một thời kỳ tăm tối trong lịch sử có liên quan đến phong trào ưu sinh. Những quốc gia phương Tây khác có lẽ sẽ áp dụng cách tiếp cận tự do hơn. Một số nước (như Trung Quốc hay Singapore, vì cả hai đều có chính sách dân số dài hạn) có thể không chỉ cho phép mà còn tích cực thúc đẩy việc áp dụng chọn lọc di truyền và kỹ thuật di truyền để tăng cường trí tuệ người dân của họ một khi đã sẵn có công nghệ để thực hiện.

Khi ví dụ đã được đưa ra và kết quả bắt đầu xuất hiện, những người do dự sẽ có động lực rất lớn để làm theo. Các dân tộc sẽ phải đối mặt với khả năng trở thành những “vùng nước tù” về nhận thức và thất bại trong các cuộc chạy đua kinh tế, khoa học, quân sự hay uy tín với các đối thủ cạnh tranh sẵn sàng đón nhận những công nghệ mới giúp tăng cường khả năng con người. Các cá nhân trong xã hội sẽ nhận ra rằng “ghế” ở các trường điểm sẽ dành cho toàn những trẻ em được chọn lọc di truyền (mà nhìn chung là xinh xắn, khỏe mạnh và chăm chỉ hơn), và sẽ muốn con cháu mình cũng có những ưu thế đó. Có khả năng thái độ xã hội sẽ dịch chuyển trong một khoảng thời gian tương đối ngắn, có thể chỉ khoảng một thập kỷ, một khi công nghệ này đã được chứng minh là hoạt động tốt và đem lại được lợi ích đáng kể. Những cuộc khảo sát ý kiến ở Mỹ đã hé lộ một sự dịch chuyển mạnh mẽ trong sự chấp thuận của công chúng đối với thụ tinh trong ống nghiệm, sau khi “đứa trẻ ống nghiệm” đầu tiên, Louise Brown, ra đời năm 1978. Vài năm trước, chỉ có 18% người Mỹ nói cá nhân họ sẽ dùng thụ tinh trong ống nghiệm để điều trị vô sinh; mặc dù vậy trong một cuộc khảo sát thực hiện không lâu sau khi Louise Brown ra đời, có đến 53% nói họ có thể sẽ làm như vậy, và con số vẫn không ngừng tăng lên.⁶¹ (Để so sánh, trong một

cuộc khảo sát thực hiện năm 2004, 28% người Mỹ chấp nhận chọn lọc phôi để có “sức mạnh hoặc trí tuệ”, 58% chấp nhận việc đó để tránh nguy cơ ung thư khi trưởng thành, và 68% chấp nhận để tránh các loại bệnh hiểm nghèo khi còn nhỏ.⁶²⁾ Nếu chúng ta tính thêm cả các khoảng trễ khác – chẳng hạn, năm đến 10 năm để thu thập thông tin cần thiết cho việc chọn lọc với hiệu quả đáng kể giữa một tập hợp phôi IVF (có thể lâu hơn nhiều trước khi các giao tử sinh ra từ tế bào gốc sẵn sàng để sử dụng cho quá trình sinh sản của con người), 10 năm để đạt mức trí tuệ đáng kể và 20 đến 25 năm để thế hệ được cải thiện đạt tới lứa tuổi mà họ bắt đầu có khả năng sinh sản – chúng ta sẽ thấy rằng các liệu pháp tăng cường dòng mầm ít có khả năng tác động đáng kể đến xã hội trước thời điểm giữa thế kỷ này. Tuy nhiên, từ thời điểm này trở đi, trí tuệ của một tỷ lệ đáng kể dân số trưởng thành có thể bắt đầu được kích thích bởi các biện pháp tăng cường di truyền. Tốc độ tiến bộ khi đó sẽ gia tăng rất nhanh vì các nhóm được thụ thai sử dụng những công nghệ di truyền thế hệ sau mạnh hơn (cụ thể là các “giao tử” sinh ra từ tế bào gốc và chọn lọc phôi lặp lại) sẽ gia nhập lực lượng lao động.

Cùng với sự phát triển đầy đủ của các công nghệ di truyền được mô tả ở trên (tạm gác lại những khả năng kỳ lạ hơn như trí tuệ trong mô thần kinh được nuôi cấy), có thể bảo đảm rằng những cá nhân mới sinh ra nhìn chung sẽ thông minh hơn bất cứ con người nào từng tồn tại, và sẽ còn không ngừng tiến bộ. Tiềm năng của các liệu pháp tăng cường sinh học do đó là vô cùng lớn, có thể đủ để đạt đến ít nhất là những dạng thức yếu của siêu trí tuệ. Đó không phải là điều đáng ngạc nhiên. Sau tất cả, các quá trình tiến hóa chậm chạp cũng đã gia tăng đáng kể trí tuệ loài người ngay cả so với những người họ

hàng gần của chúng ta là loài linh trưởng lớn và tổ tiên vượn nhân hình. Chúng ta cũng không có lý do gì để cho rằng các Homo sapiens đã đạt đến đỉnh cao của hiệu quả nhận thức có thể đạt được trong một hệ sinh học. Chúng ta không phải là loài sinh vật thông minh nhất, mà thậm chí còn nên tự nghĩ rằng mình là loài ấu trĩ nhất có thể khởi động một nền văn minh công nghệ – một khe hẹp mà chúng ta chiếm được chỉ vì đến trước chứ không phải vì thích nghi tốt với nó.

Tiến bộ trên con đường sinh học rõ ràng là khả thi. Độ trễ thể hệ trong liệu pháp can thiệp dòng mầm đồng nghĩa với việc tiến bộ có thể không bất ngờ và đột ngột như trong những kịch bản liên quan đến trí tuệ máy. (Các liệu pháp gen sinh dưỡng và can thiệp dược lý trên lý thuyết có thể bỏ qua độ trễ thể hệ và ít có khả năng tạo ra những hiệu ứng mạnh). Tiềm năng tối thượng của trí tuệ máy tất nhiên sẽ cao hơn trí tuệ hữu cơ. (Ta có thể cảm nhận được độ lớn của khoảng cách này bằng cách xem xét sự khác biệt về tốc độ giữa các linh kiện điện tử và tế bào thần kinh: các bóng bán dẫn ngày nay cũng hoạt động trên thang thời gian ngắn hơn các neuron sinh học hàng chục triệu lần). Tuy nhiên, ngay cả những biện pháp tăng cường tương đối khiêm tốn của nhận thức sinh học cũng có thể có hệ quả quan trọng. Cụ thể, tăng cường nhận thức có thể gia tăng tốc độ của khoa học và công nghệ, bao gồm cả tiến bộ hướng tới những hình thức khuếch đại trí tuệ sinh học và trí tuệ máy tiềm năng hơn. Hãy xem tốc độ tiến bộ trong lĩnh vực trí tuệ nhân tạo có thể thay đổi như thế nào trong một thế giới nơi mà một chàng Joe Tầm thường lại là người có trí tuệ ngang với Alan Turing hay John von Neuman,

và trong đó có hàng triệu người vượt qua được những bộ óc vĩ đại trong quá khứ.⁶³

Chúng ta hãy thảo luận về những tác động chiến lược của việc tăng cường nhận thức trong một chương khác ở dưới, nhưng có thể tóm tắt phần này bằng cách ghi nhận ba kết luận: (1) siêu trí tuệ, ít nhất là ở những dạng thức yếu, có thể đạt được bằng các biện pháp tăng cường mang tính sinh học-công nghệ; (2) tính khả thi của người được tăng cường nhận thức sẽ làm gia tăng khả năng những hình thức tiên tiến hơn của trí tuệ máy sẽ trở nên khả thi – vì ngay cả nếu chúng ta về cơ bản không thể tạo ra trí tuệ máy (mặc dù không có lý do gì để giả định như vậy), thì trí tuệ máy vẫn có thể nằm trong tầm với của những người được tăng cường nhận thức; và (3) khi xem xét những kịch bản kéo dài đến nửa sau của thế kỷ này và xa hơn nữa, chúng ta cần phải tính đến sự xuất hiện khả thể của một thế hệ dân số được tăng cường về mặt di truyền – các cử tri, nhà sáng chế, nhà khoa học – với mức độ tăng cường được nâng cao nhanh chóng trong những thập kỷ tiếp theo.

Giao diện não-máy tính

Đôi khi có ý kiến cho rằng các giao diện não-máy tính trực tiếp, cụ thể là cấy ghép, có thể giúp con người khai thác được sức mạnh của tính toán kỹ thuật số – trí nhớ hoàn hảo, tính toán số học nhanh chóng và chính xác, và truyền dữ liệu bằng thông cao – cho phép hệ thống lai này vượt trội về cơ bản so với bộ não không được tăng cường.⁶⁴ Nhưng cho dù khả năng kết nối trực tiếp giữa não bộ con người và máy tính đã được chứng minh, thì những giao diện như vậy cũng khó có thể được dùng làm biện pháp tăng cường trong tương lai gần.⁶⁵

Trước hết, có những rủi ro đáng kể về biến chứng y học – nhiễm trùng, điện cực lạc vị, xuất huyết và suy giảm nhận thức – khi cấy điện cực vào não. Có lẽ minh họa rõ ràng nhất cho tới nay về những lợi ích có thể có được nhờ kích thích não là việc chữa trị cho bệnh nhân Parkinson. Bộ phận cấy ghép cho bệnh nhân Parkinson tương đối đơn giản: nó không thực sự liên hệ với não bộ mà chỉ đơn giản cung cấp một dòng điện kích thích đến nhân dưới đồi. Một video minh họa đã chiếu cảnh bệnh nhân đang ngã người trên ghế và hoàn toàn bất động bất ngờ trở nên tươi tỉnh khi dòng điện được bật lên: người này có thể cử động tay, đứng dậy đi quanh phòng, quay lại và thực hiện một bước nhảy xoay vòng trên đầu ngón chân. Mặc dù vậy, phía sau thủ thuật đặc biệt đơn giản và thành công màu nhiệm này vẫn tồn tại những nhược điểm. Một nghiên cứu cho thấy nhóm các bệnh nhân Parkinson có bộ phận cấy ghép nằm sâu trong não có sự suy giảm về tính lưu loát của lời nói, sự chú ý chọn lọc, khả năng gọi tên màu sắc và ký ức bằng lời so với nhóm đối chứng. Nghiên cứu cũng ghi nhận rằng những đối tượng được điều trị than phiền nhiều hơn về nhận thức.⁶⁶ Những rủi ro và tác dụng phụ này có thể chấp nhận được nếu thủ thuật được dùng để giảm bớt tình trạng tàn phế nặng nề. Nhưng để những đối tượng khỏe mạnh tình nguyện được phẫu thuật thần kinh thì tình trạng hoạt động bình thường cần phải được tăng cường đáng kể.

Điều này đưa chúng ta đến lý do thứ hai để nghi ngờ việc siêu trí tuệ có thể đạt được nhờ chuyển hóa sang dạng nửa người nửa máy (cyborgization), nghĩa là tăng cường có lẽ sẽ khó hơn rất nhiều so với trị liệu. Những bệnh nhân bại liệt có thể có được lợi ích từ một bộ phận cấy ghép thay thế cho những dây thần kinh bị hỏng của họ

hoặc kích hoạt các bộ tạo sinh mẫu hình vận động của cột sống.⁶⁷ Những bệnh nhân khiếm thính hoặc khiếm thị có thể hưởng lợi từ ốc tai hay võng mạc nhân tạo.⁶⁸ Người mắc bệnh Parkinson hoặc gặp phải các cơn đau kinh niên có thể hưởng lợi từ việc kích thích sâu trong não để kích hoạt hoặc ức chế hoạt động trong một khu vực cụ thể của não bộ.⁶⁹ Điều xem ra khó đạt được hơn rất nhiều là tương tác bằng thông cao trực tiếp giữa não bộ và máy tính để gia tăng đáng kể trí tuệ ở một dạng thức không dễ dàng có được thông qua các phương pháp khác. Hầu hết những lợi ích tiềm ẩn mà cấy ghép não có thể cung cấp cho những đối tượng khỏe mạnh đều có thể có được với rủi ro, chi phí và sự bất tiện ít hơn rất nhiều bằng cách sử dụng các cơ quan vận động và cảm giác của chúng ta để tương tác với những chiếc máy tính nằm ngoài cơ thể mình. Chúng ta không cần cắm cáp quang vào não bộ để truy cập Internet. Võng mạc người không chỉ có thể truyền dữ liệu với tốc độ ấn tượng 10 triệu bit/giây mà còn được đóng gói sẵn với một lượng lớn “phần ẩm” (wetware) chuyên dụng, là vỏ não thị giác, vốn đã được điều chỉnh rất phù hợp để trích xuất ý nghĩa từ dòng thông tin này và giao tiếp với các khu vực não bộ khác để tiếp tục xử lý.⁷⁰ Thậm chí nếu có một cách thức dễ dàng khác để “bơm” thêm thông tin vào não bộ thì luồng dữ liệu bổ sung cũng không làm được gì nhiều để nâng cao tốc độ suy nghĩ và học tập của chúng ta, trừ khi toàn bộ bộ máy thần kinh cần thiết để trích xuất ý nghĩa của dữ liệu được nâng cấp một cách tương tự. Vì nó bao gồm hầu hết não bộ nên thứ cần phải có là một “bộ não giả hoàn chỉnh” – đây chỉ là một cách khác để nói đến trí tuệ nhân tạo tổng thể. Nhưng nếu có một AI cấp độ con người, ta có thể bỏ qua phẫu thuật thần kinh: một máy tính cũng có thể có vỏ máy kim loại

giống như người có xương. Vậy trường hợp giới hạn này chỉ đưa chúng ta quay lại con đường AI mà mình vừa xem xét.

Giao diện não-máy tính cũng được đề xuất như một phương pháp lấy thông tin từ não bộ nhằm mục đích giao tiếp với các bộ não khác hoặc với máy móc.⁷¹ Những liên kết ngược (uplink) này giúp các bệnh nhân mắc hội chứng khóa trong (locked-in syndrome) giao tiếp với thế giới bên ngoài bằng cách dịch chuyển con trỏ trên màn hình bằng ý nghĩ.⁷² Bảng thông có được trong những thí nghiệm này là thấp: bệnh nhân vất vả, chậm chạp đánh máy từng chữ với tốc độ vài từ một phút. Ta có thể dễ dàng hình dung ra các phiên bản cải tiến của công nghệ này – một bộ phận cấy ghép thể hệ sau có thể được cấy vào vùng Broca (khu vực trong thùy trán tham gia tạo sản ngôn ngữ) và thu thập lời nói bên trong.⁷³ Nhưng trong khi công nghệ như vậy có thể trợ giúp một vài người bị tàn tật vì đột quỵ hay thoái hóa cơ, thì nó lại ít hấp dẫn đối với những người khỏe mạnh. Chức năng do nó cung cấp về cơ bản là một microphone gắn với phần mềm nhận dạng tiếng nói, hiện đã được thương mại hóa rộng rãi – loại bỏ được những cơn đau, sự bất tiện, chi phí và những rủi ro gắn với phẫu thuật thần kinh (và cũng loại bỏ ít nhất một vài ẩn ý về thiết bị nghe trộm gắn trong sọ não). Cất giữ máy móc của chúng ta bên ngoài cơ thể cũng sẽ giúp nâng cấp chúng dễ dàng hơn.

Nhưng còn giấc mơ về việc hoàn toàn bỏ qua ngôn từ và thiết lập một kết nối giữa hai bộ não cho phép “tải về” các khái niệm, ý nghĩ hay toàn bộ một lĩnh vực chuyên môn từ bộ não này sang bộ não khác thì sao? Chúng ta có thể tải những tệp tin lớn về máy tính, bao gồm những thư viện với hàng triệu cuốn sách và bài báo chỉ trong vòng vài giây: vậy ta có thể làm tương tự với não bộ của mình hay

không? Tính khả thi rõ ràng của ý tưởng này có thể đến từ một quan điểm không chính xác về cách thức thông tin được lưu trữ và biểu diễn trong não bộ. Như đã lưu ý, rào chắn hạn chế tốc độ trong trí tuệ con người không phải là việc dữ liệu được đưa vào não nhanh đến mức nào mà là não bộ có thể chiết xuất được ý nghĩa và hiểu được dữ liệu nhanh đến đâu. Có lẽ, chúng ta nên truyền tải trực tiếp ý nghĩa thay vì đóng gói chúng vào các dữ liệu giác quan mà người nhận phải giải mã. Có hai vấn đề với điều này. Thứ nhất là não bộ, trái với những loại chương trình mà chúng ta chạy trong máy tính, không sử dụng các định dạng lưu trữ và biểu diễn dữ liệu đã chuẩn hóa. Thay vào đó, mỗi bộ não phát triển những cách biểu diễn đặc trưng của mình cho những nội dung cấp cao hơn. Việc những tổ hợp neuron cụ thể nào được sử dụng để biểu diễn một khái niệm nhất định phụ thuộc vào những trải nghiệm đặc thù của bộ não đang được xem xét (cùng với một số tác nhân di truyền và các quá trình sinh lý ngẫu nhiên). Cũng như trong các mạng lưới thần kinh nhân tạo, ý nghĩa trong mạng lưới thần kinh sinh học có lẽ được biểu diễn một cách toàn diện trong những mẫu hình cấu trúc và hoạt động của những khu vực chồng lấn lớn, mà không phải trong những tế bào bộ nhớ riêng biệt được sắp xếp thành mạng lưới gọn gàng.⁷⁴ Do đó, không thể thiết lập một ánh xạ đơn giản giữa các neuron trong một bộ não với các neuron trong một bộ não khác, sao cho suy nghĩ có thể tự động truyền qua. Để ý nghĩ của một bộ não có thể hiểu được bởi một bộ não khác, ý nghĩ này cần phải được phân rã và đóng gói thành một hệ thống ký hiệu theo một quy ước chung nào đó, cho phép ký hiệu có thể được bộ não nhận diện giải mã đúng. Đó chính là nhiệm vụ của ngôn ngữ.

Về nguyên tắc, ta có thể hình dung ra việc giảm tải nhận thức của việc khởi tạo và diễn giải sang một giao diện có thể bằng cách nào đó đọc được những trạng thái neuron trong não bộ người gửi và đưa một mẫu hình kích hoạt có trước vào não người nhận. Nhưng điều này cũng đưa chúng ta đến vấn đề thứ hai với kịch bản nửa người nửa máy. Ngay cả khi bỏ sang một bên thách thức kỹ thuật (khá khổng lồ) của việc làm thế nào để cùng lúc đọc và ghi một cách đáng tin cậy từ hàng tỷ neuron có thể được đánh địa chỉ, thì việc tạo ra giao diện cần thiết vẫn có thể là một vấn đề AI toàn diện. Giao diện này sẽ cần có một cấu phần có thể ánh xạ (theo thời gian thực) những mẫu hình kích hoạt trong một bộ não vào những mẫu hình kích hoạt tương đương về mặt ngữ nghĩa trong bộ não kia. Hiểu biết chi tiết nhiều cấp độ về điện toán thần kinh cần để thực hiện tác vụ này xem ra sẽ trực tiếp tạo ra AI hình thái thần kinh.

Ngay cả với những hạn chế này, con đường người lai máy dẫn đến việc tăng cường nhận thức vẫn không phải là hoàn toàn bế tắc. Công trình ấn tượng trên hồi hải mã của chuột đã cho thấy tính khả thi của một bộ phận neuron giả có thể tăng cường hiệu năng trong một tác vụ bộ nhớ làm việc đơn giản.⁷⁵ Trong phiên bản hiện tại, bộ phận cấy ghép thu thập thông tin đầu vào từ 12-24 điện cực đặt trong một khu vực (“CA3”) của hồi hải mã và phóng chiếu nó vào một số lượng neuron tương tự trong một khu vực khác (“CA1”). Một bộ vi xử lý được huấn luyện để phân biệt giữa hai mẫu hình kích hoạt khác nhau trong khu vực thứ nhất (tương ứng với hai ký ức khác nhau, “cần gạt phải” hoặc “cần gạt trái”), và để học cách những mẫu hình này được phóng chiếu sang khu vực thứ hai. Bộ phận giả này không chỉ có thể khôi phục chức năng khi kết nối neuron bình thường giữa

hai vùng neuron bị chặn, mà còn có thể tăng cường hiệu năng của tác vụ ghi nhớ tới một mức độ mà chuột thường không thể có được bằng cách gửi một mã thông báo đặc biệt rõ ràng của một mẫu hình ký ức cụ thể tới khu vực thứ hai. Mặc dù đã là một thành tựu kỹ thuật theo những tiêu chuẩn hiện đại, nghiên cứu này vẫn để lại nhiều câu hỏi khó chưa được trả lời: cách tiếp cận này có hiệu quả đến đâu khi số lượng ký ức lớn hơn? Chúng ta có thể kiểm soát đến mức nào sự bùng nổ tổ hợp mà, nếu không được kiểm soát, có thể khiến cho việc học được ánh xạ đúng trở nên không khả thi khi số lượng neuron đầu vào và đầu ra tăng lên? Liệu hiệu năng được tăng cường trong tác vụ thử nghiệm có đi kèm với một cái giá tiềm tàng nào đó hay không, ví dụ như suy giảm khả năng khái quát hóa từ kích thích cụ thể được sử dụng trong thí nghiệm, hoặc suy giảm khả năng loại bỏ (unlearn) mối liên hợp khi môi trường thay đổi? Liệu các đối tượng thí nghiệm có được hưởng lợi theo một cách nào đó, ngay cả nếu – khác với chuột – họ có thể sử dụng các loại công cụ trợ giúp ghi nhớ bên ngoài như bút và giấy? Và việc áp dụng phương pháp tương tự vào những phần khác của não bộ còn khó khăn hơn bao nhiêu? Trong khi bộ phận giả hiện tại tận dụng cấu trúc truyền thẳng tương đối đơn giản của các phần của hồi hải mã (về cơ bản phục vụ như một cầu nối một chiều giữa các vùng CA3 và CA1), các cấu trúc khác trong vỏ não có những vòng phản hồi quấn vào nhau làm tăng đáng kể độ phức tạp của sơ đồ móc nối (wiring) và có thể là cả độ khó của việc giải mã chức năng của bất cứ nhóm neuron được nhúng nào.

Một niềm hy vọng của con đường nửa người nửa máy là não bộ, nếu được cấy ghép vĩnh viễn một thiết bị kết nối nó với nguồn bên ngoài, theo thời gian có thể học được một cách ánh xạ hiệu quả giữa

các trạng thái nhận thức nội tại của nó và những thông tin đầu vào nó nhận được từ thiết bị, hoặc những thông tin đầu ra mà thiết bị nhận được. Khi đó, bộ phận cấy ghép không cần thông minh, mà não bộ sẽ thích ứng một cách thông minh với giao diện. Điều này rất giống với cách não bộ của một đứa trẻ sơ sinh dần học được cách giải thích những tín hiệu đến từ các thụ thể trong mắt và tai của mình.⁷⁶ Nhưng một lần nữa ta phải đặt câu hỏi rằng thực sự sẽ có bao nhiêu lợi ích. Giả sử não bộ có thể linh hoạt tới mức phát hiện được ra mẫu hình trong một số luồng thông tin đầu vào mới, ngẫu nhiên được phóng chiếu vào phần nào đó của vỏ não bằng giao diện não-máy tính, thì tại sao lại không phóng chiếu thông tin đó lên võng mạc, như một mẫu hình thị giác, hoặc lên ốc tai dưới dạng âm thanh? Phương án thay thế kiểu “low-tech” này tránh được hàng ngàn biến chứng, và trong cả hai trường hợp, não bộ đều có thể triển khai các cơ chế nhận dạng mẫu hình và tính linh hoạt để học cách hiểu được thông tin.

Các mạng lưới và tổ chức

Một con đường khác dẫn đến siêu trí tuệ có thể hình dung ra được là thông qua việc dần mở rộng các mạng lưới và tổ chức liên kết trí não của những con người riêng lẻ, cũng như với các robot và bộ phận giả. Ý tưởng không phải là việc này sẽ tăng cường năng lực trí tuệ của các cá nhân đủ để biến họ thành siêu trí tuệ, mà là một hệ thống nào đó được hình thành từ các cá nhân được kết nối và tổ chức theo cách này có thể có được một hình thức siêu trí tuệ, thứ mà trong chương sau chúng tôi sẽ mô tả kỹ lưỡng hơn với tên gọi “siêu trí tuệ tập thể”.⁷⁷

Loài người đã đạt được nhiều thành công vĩ đại về trí tuệ tập thể trong suốt chiều dài lịch sử. Những kết quả này đến từ nhiều nguồn, bao gồm những sáng tạo trong công nghệ truyền thông như viết và in ấn, và trên hết là việc sử dụng ngôn ngữ; sự tăng trưởng về số lượng và mật độ dân số trên thế giới; những cải tiến khác nhau trong các kỹ thuật tổ chức và tiêu chuẩn trí thức; và sự tích lũy dần dần nguồn vốn thể chế. Nói chung, trí tuệ tập thể của một hệ thống bị giới hạn bởi năng lực của các trí óc thành viên trong hệ thống, những chi phí trong việc truyền tải thông tin liên quan giữa họ và những biến dạng và tình trạng kém hiệu năng đang tràn ngập trong các tổ chức của con người. Nếu chi phí truyền thông giảm (không chỉ chi phí thiết bị mà cả các khoảng trễ phản hồi, các gánh nặng về thời gian và sự chú ý và các tác nhân khác nữa) thì các tổ chức lớn hơn và được kết nối với mật độ cao hơn sẽ trở nên khả thi. Điều này cũng có thể xảy ra nếu tìm được các biện pháp sửa chữa đối với sự méo mó quan liêu làm biến dạng cuộc sống tổ chức – những trò chơi địa vị đầy lãng phí, sự sai lệch sứ mệnh, sự giấu diếm hoặc làm giả thông tin và những vấn đề đại diện khác. Ngay cả các giải pháp không toàn diện cho những vấn đề này cũng có thể đem lại lợi ích lớn cho trí tuệ tập thể.

Những đổi mới sáng tạo về công nghệ và thể chế có thể đóng góp vào quá trình phát triển trí tuệ tập thể của chúng ta là rất nhiều và đa dạng. Ví dụ, các thị trường dự đoán được trợ giá có thể nuôi dưỡng các chuẩn mực tìm kiếm sự thực và cải thiện dự báo về các vấn đề khoa học và xã hội đang gây tranh cãi.⁷⁸ Máy phát hiện nói dối (nếu có thể chế tạo được những chiếc máy đáng tin cậy và dễ sử dụng) có thể giảm mức độ gian dối trong các hoạt động của con

người.⁷⁹ Máy phát hiện nói dối có thể còn mạnh hơn nữa.⁸⁰ Thậm chí nếu không có các công nghệ trí não mới lạ thì một vài hình thức dối lừa có thể trở nên khó thực hiện hơn nhờ vào sự sẵn có của nhiều loại dữ liệu, bao gồm uy tín và thành tích, hoặc việc ban hành những chuẩn mực trí thức và một nền văn hóa duy lý. Việc giám sát tự nguyện hoặc không tự nguyện sẽ tạo ra khối lượng thông tin khổng lồ về hành vi con người. Các trang mạng xã hội đã được hàng tỷ người sử dụng để chia sẻ các chi tiết cá nhân: không lâu nữa những người này có thể bắt đầu tải lên các bản ghi liên tục về cuộc sống của mình từ các micro và máy quay video tích hợp trong điện thoại di động hoặc gọng kính. Việc tự động phân tích những luồng thông tin này có thể tạo ra nhiều ứng dụng mới (tất nhiên là cả tích cực và tiêu cực).⁸¹

Tăng trưởng trong trí tuệ tập thể còn có thể tới từ những cải thiện tổ chức và kinh tế tổng thể hơn, và từ việc tăng tỷ lệ dân số thế giới được giáo dục, kết nối số và tích hợp vào nền văn hóa trí tuệ toàn cầu.⁸²

Internet nổi lên như một ranh giới đặc biệt năng động cho những sáng tạo và thử nghiệm. Hầu hết tiềm năng của nó vẫn chưa được khai thác hết. Sự phát triển liên tục của một trang web thông minh hỗ trợ tốt hơn cho việc tranh luận, loại bỏ thiên kiến và tổng hợp ý kiến có thể đóng góp rất nhiều vào việc tăng cường trí tuệ tập thể của loài người như một khối tổng thể, cũng như của các nhóm cụ thể.

Nhưng còn ý tưởng xem ra huyền ảo hơn cho rằng Internet có thể “thức giấc” một ngày đẹp trời nào đó? Liệu Internet có thể trở thành thứ gì đó khác, hơn là chỉ làm xương sống cho một siêu trí tuệ tập thể được tích hợp lỏng lẻo – giống như một hộp sọ ảo đựng một

thứ siêu trí thức hợp nhất đang nổi lên? (Đây là một trong những cách mà trí tuệ nhân tạo có thể trỗi dậy, theo bài luận có tầm ảnh hưởng của Vernor Vinge năm 1993, trong đó đưa ra thuật ngữ “điểm kỳ dị công nghệ”.⁸³) Ta có thể lý luận chống lại điều này rằng trí tuệ máy đã đủ khó để làm ra được bằng kỹ thuật phức tạp, và thật là đáng kinh ngạc khi giả định rằng nó sẽ trỗi dậy một cách tự phát. Tuy nhiên, câu chuyện không nhất thiết phải là một phiên bản tương lai nào đó của Internet bỗng nhiên trở thành siêu trí tuệ hoàn toàn ngẫu nhiên. Một phiên bản khả dĩ hơn của kịch bản này có thể là Internet tích tụ các cải tiến thông qua công sức của nhiều người trong nhiều năm – các công trình để làm ra các thuật toán tìm kiếm và lọc thông tin tốt hơn, các định dạng biểu diễn dữ liệu mạnh hơn, các tác tử phần mềm tự chủ có năng lực cao hơn cùng các giao thức hiệu quả hơn điều chỉnh việc tương tác giữa các “bot” như vậy – và vô số những cải tiến gia tăng này cuối cùng sẽ tạo ra cơ sở cho một dạng thức hợp nhất cao hơn nào đó của trí tuệ mạng. Ít nhất thì có vẻ như một hệ thống nhận thức trên nền web như vậy (đã siêu bão hòa với công suất máy tính và những nguồn tài nguyên khác cần thiết cho tăng trưởng bùng nổ trừ một thành phần thiết yếu nào đó) có thể, khi thành tố cuối cùng này được “ném” vào chảo, sẽ bùng lên thành siêu trí tuệ. Dù sao đi nữa, loại kịch bản này hội tụ thành một con đường khả thể khác dẫn đến siêu trí tuệ, con đường trí tuệ nhân tạo tổng thể mà chúng ta đã thảo luận.

Tóm tắt

Sự thật rằng có nhiều con đường dẫn đến siêu trí tuệ sẽ làm tăng sự tự tin của chính chúng ta rằng cuối cùng mình cũng sẽ đến đích.

Nếu một con đường nào đó bị chặn, chúng ta vẫn có thể tiếp tục tiến lên.

Nhiều con đường không có nghĩa là nhiều điểm đến. Thậm chí nếu như sự khuếch đại trí tuệ đáng kể có thể đạt được lần đầu tiên với một trong những con đường trí tuệ phi máy móc, thì điều đó cũng không làm cho trí tuệ máy trở nên không phù hợp. Ngược lại, trí tuệ sinh học hay tổ chức được tăng cường có thể đẩy nhanh sự phát triển của khoa học công nghệ, và có khả năng làm cho các dạng thức khuếch đại trí tuệ toàn diện hơn như giả lập hoàn chỉnh não bộ và AI tới đích nhanh hơn.

Tuy nhiên, điều này không đồng nghĩa với việc cách thức đến với siêu trí tuệ máy là không quan trọng. Đường tới đích có thể tạo ra khác biệt lớn cho kết quả cuối cùng. Thậm chí nếu những năng lực tối hậu mà cuối cùng chúng ta có được không phụ thuộc nhiều vào con đường thì cách sử dụng những năng lực này – hay mức độ mà loài người có thể kiểm soát “tâm tính” của chúng – có thể phụ thuộc nhiều vào những chi tiết trong cách tiếp cận của chúng ta. Ví dụ, những sự tăng cường cho trí tuệ sinh học hoặc trí tuệ tổ chức có thể làm tăng khả năng dự báo rủi ro của chúng ta cũng như khả năng thiết kế ra siêu trí tuệ máy an toàn và hữu ích. (Một đánh giá chiến lược đầy đủ sẽ liên quan đến nhiều vấn đề phức tạp, và ta sẽ bàn đến trong Chương 14.)

Siêu trí tuệ thực sự (trái với những gia tăng nhỏ trong các cấp độ trí tuệ hiện tại) có thể đạt được trước nhất qua con đường AI. Tuy nhiên, trên con đường này cũng có nhiều sự bất định cơ bản. Điều này khiến việc đánh giá một cách thận trọng con đường này dài bao nhiêu hay có bao nhiêu chướng ngại vật trở nên khó khăn. Con

đường giả lập hoàn chỉnh não bộ cũng có đôi chút khả năng để trở thành tuyến đường nhanh nhất dẫn tới siêu trí tuệ. Con đường này cơ bản chỉ đòi hỏi những tiến bộ công nghệ gia tăng thay vì những đột phá lý thuyết, vì thế ta có thể có lý do để tin rằng cuối cùng nó cũng sẽ thành công. Tuy nhiên, xem ra có xác suất tương đối cao rằng nếu tiến triển trên con đường giả lập hoàn chỉnh não bộ có nhanh đi chẳng nữa thì trí tuệ nhân tạo vẫn sẽ tới đích trước: đó là do tính khả thi của các AI hình thái học thần kinh dựa trên giả lập từng phần.

Tăng cường nhận thức sinh học rõ ràng là khả thi, cụ thể là những tăng cường dựa trên chọn lọc di truyền. Chọn lọc phôi lặp lại hiện đang được coi là một công nghệ đặc biệt hứa hẹn. Tuy nhiên, so với những đột phá khả thể trong trí tuệ máy, những biện pháp tăng cường sinh học có thể tương đối chậm chạp và lẫn mờ. Khả quan nhất, chúng cũng chỉ có thể tạo ra những dạng thức yếu của siêu trí tuệ mà thôi (chúng ta sẽ đề cập ngay đến vấn đề này).

Tính khả thi rõ ràng của tăng cường sinh học có thể làm tăng sự tự tin của chúng ta rằng trí tuệ máy cuối cùng rồi cũng được tạo ra, vì các nhà khoa học và kỹ sư được tăng cường sẽ có thể đạt tiến độ nhanh và nhiều hơn so với các đồng nghiệp phát triển tự nhiên của họ. Đặc biệt trong những kịch bản mà ở đó trí tuệ máy bị trì hoãn sang quá nửa thế kỷ sau, những tập thể người được tăng cường về nhận thức dần xuất hiện trên vũ đài sẽ đóng vai trò quan trọng hơn trong những phát triển tiếp theo.

Giao diện não-máy tính xem ra khó có thể trở thành một nguồn phát sinh siêu trí tuệ. Những cải tiến trong mạng lưới và tổ chức trong dài hạn có thể tạo ra những dạng thức siêu trí tuệ yếu của trí

tuệ tập thể; nhưng hẳn là chúng sẽ đóng vai trò hỗ trợ tương tự như biện pháp tăng cường nhận thức sinh học, nâng cao dần khả năng giải quyết hiệu quả đối với các vấn đề trí tuệ của con người. So với những biện pháp tăng cường sinh học, các tiến bộ trong mạng lưới và tổ chức sẽ tạo ra khác biệt sớm hơn – thực tế, những tiến bộ này xảy ra liên tục và đã có tác động đáng kể. Tuy nhiên những cải tiến trong mạng lưới và tổ chức có thể làm gia tăng khả năng giải quyết vấn đề của chúng ta trong phạm vi hẹp hơn so với những cải tiến trong nhận thức sinh học – thúc đẩy “trí tuệ tập thể” thay vì “trí tuệ chất lượng”, để chờ đón một sự khác biệt mà chúng tôi sẽ giới thiệu trong chương sau.

Chương 3

Các dạng thức siêu trí tuệ

Vậy chính xác thì ý chúng tôi là gì khi nói về “siêu trí tuệ”? Vì không mong muốn sa vào đầm lầy thuật ngữ nên chúng ta cần phải nói gì đó để làm rõ nền tảng khái niệm.

Chương này xác định ba hình thức khác nhau của siêu trí tuệ và lập luận rằng theo một lớp nghĩa phù hợp với thực tế thì chúng tương đương nhau. Chúng tôi cũng cho thấy rằng tiềm năng của trí tuệ trên chất nền (substrate) máy móc lớn hơn nhiều so với trên chất nền sinh học. Máy móc có một số ưu thế cơ bản cho chúng sự áp đảo. Con người sinh học, cho dù có được tăng cường, rồi cũng sẽ bị vượt qua.

Nhiều loại máy móc và động vật không phải con người đã hoạt động ở các cấp độ siêu nhân trong những phạm vi hẹp. Dơi giải nghĩa tín hiệu siêu âm tốt hơn người, máy tính giỏi số học hơn chúng ta và các chương trình chơi cờ đánh thắng chúng ta trong môn cờ vua. Danh sách những tác vụ cụ thể có thể được phần mềm thực thi tốt hơn sẽ còn tiếp tục kéo dài. Nhưng cho dù các hệ thống xử lý thông tin chuyên biệt sẽ có nhiều ứng dụng, vẫn còn những vấn đề cơ bản chỉ xuất hiện cùng với viễn cảnh rằng trí tuệ máy có đủ trí tuệ tổng quát để thay thế con người trong mọi lĩnh vực.

Như đã đề cập đến ở trên, chúng tôi sử dụng thuật ngữ “siêu trí tuệ” để nói đến những trí tuệ vượt trội hơn các bộ óc con người thông minh nhất hiện nay rất nhiều, xuyên suốt nhiều lĩnh vực nhận thức rất phổ thông. Thuật ngữ này vẫn còn khá mờ lung. Các loại hệ thống khác nhau với các thuộc tính hiệu năng tách biệt có thể được xếp loại là siêu trí tuệ theo định nghĩa này. Để tiếp tục phân tích, sẽ hữu ích nếu ta phân tách định nghĩa đơn giản này của siêu trí tuệ bằng cách phân biệt những tổ hợp siêu năng lực trí tuệ khác nhau. Có nhiều cách để thực hiện việc phân tách này. Ở đây, chúng tôi sẽ phân biệt giữa ba dạng thức: siêu trí tuệ tốc độ, siêu trí tuệ tập thể và siêu trí tuệ chất lượng.

Siêu trí tuệ tốc độ

Siêu trí tuệ tốc độ cũng giống như tâm trí của con người nhưng nhanh hơn rất nhiều. Về mặt khái niệm, đây là dạng thức siêu trí tuệ dễ phân tích nhất.¹

Siêu trí tuệ tốc độ: Một hệ thống có thể làm mọi thứ mà một người trí thức có thể làm được, nhưng nhanh hơn nhiều.

Khi nói về “nhiều”, chúng tôi muốn nhắc đến một thứ gì đó giống như “nhiều cấp khuếch đại”. Nhưng thay vì cố gắng loại bỏ hoàn toàn mọi tàn dư của sự mờ lung ra khỏi định nghĩa, chúng tôi sẽ uỷ thác cho bạn đọc để giải thích nó một cách hợp lý.²

Ví dụ đơn giản nhất của siêu trí tuệ tốc độ có thể là một bản giả lập hoàn chỉnh não bộ chạy trên một phần cứng tốc độ cao.³ Một bản giả lập vận hành với tốc độ nhanh hơn bộ não sinh học hàng chục ngàn lần có thể đọc một cuốn sách trong vài giây hay viết một luận án tiến sĩ chỉ trong một buổi chiều. Với hệ số tốc độ bằng một triệu,

một bản giả lập có thể hoàn thành toàn bộ công trình trí tuệ của cả thiên niên kỷ chỉ trong vòng một ngày làm việc.⁴

Đối với một trí óc nhanh như vậy, những sự kiện trong thế giới bên ngoài có vẻ như chỉ diễn ra như một thước phim quay chậm. Giả sử, trí óc của bạn chạy nhanh hơn 10.000 lần. Nếu một người bạn bình thường nào đó đánh rơi chén trà, bạn có thể ngắm chiếc chén sứ từ từ rơi xuống thảm trong vài giờ đồng hồ, như một ngôi sao chổi lạng lẹ lướt qua không gian đến một hành tinh xa xôi; và khi dự báo về một cú va đập sắp tới chậm chạp lan truyền qua những nếp nhăn trong phần chất xám của anh bạn kia và từ đó tới hệ thần kinh ngoại biên, bạn có thể quan sát được cơ thể anh ta dần dần dẫn vì bất ngờ trong khoảng thời gian đủ dài để bạn không chỉ gọi một chén mới mà còn đọc được một hai bài báo khoa học và đánh một giấc ngon lành.

Chính vì sự giãn nở thời gian rõ ràng này của thế giới vật chất, một siêu trí tuệ tốc độ sẽ thích làm việc với các đối tượng kỹ thuật số. Nó có thể sống trong thực tại ảo và giao dịch trong nền kinh tế thông tin. Hoặc nó cũng có thể tương tác với môi trường vật lý bằng những bộ thao tác có kích thước nano, vì “chân tay” với kích thước nhỏ như vậy có thể vận hành nhanh hơn nhiều so với các phần phụ kích thước lớn. (Tần số đặc trưng của một hệ thống thường tỷ lệ nghịch với chiều dài của nó.⁵) Một trí óc nhanh nhẹn có thể hợp tác chủ yếu với các trí óc nhanh nhẹn khác thay vì với những con người chậm chạp, ì ạch.

Tốc độ ánh sáng ngày càng trở thành một cản trở quan trọng hơn khi trí óc nhanh hơn, vì những trí óc nhanh hơn đối diện với các chi phí cơ hội lớn hơn khi dùng thời gian của mình để đi lại hay giao tiếp qua những khoảng cách xa.⁶ Ánh sáng nhanh hơn máy bay phản lực

khoảng một triệu lần, nên một tác tử kỹ thuật số với tốc độ tư duy nhanh hơn một triệu lần đi vòng quanh thế giới sẽ phải tốn khoảng thời gian chủ quan tương đương với một người đi du lịch cùng thời. Gọi điện thoại đường dài có thể cũng mất thời gian như đến gặp “trực tiếp”, mặc dù rẻ hơn vì một cuộc gọi cần ít băng thông hơn. Những tác tử với tốc độ tư duy lớn muốn thường xuyên trao đổi sẽ thấy rằng “xích lại gần nhau” là có lợi. Những trí óc siêu nhanh có nhu cầu tương tác thường xuyên (chẳng hạn như thành viên của một nhóm làm việc) có thể nằm ở trong những máy tính đặt trong cùng một tòa nhà để tránh những sự chậm trễ gây khó chịu.

Siêu trí tuệ tập thể

Một dạng thức khác của siêu trí tuệ là một hệ thống có được hiệu năng vượt trội bằng cách tập hợp số lượng lớn các trí tuệ nhỏ hơn:

Siêu trí tuệ tập thể: Một hệ thống hình thành từ một số lượng lớn các trí tuệ nhỏ hơn sao cho hiệu năng chung của hệ thống trong nhiều lĩnh vực rất phổ thông vượt trội hơn rất nhiều so với hiệu năng của bất cứ hệ thống nhận thức hiện tại nào.

Siêu trí tuệ tập thể ít rõ ràng hơn về mặt khái niệm so với siêu trí tuệ tốc độ.⁷ Mặc dù vậy, nó quen thuộc hơn về mặt thực nghiệm. Trong khi không có kinh nghiệm với những trí óc cấp độ con người nhưng có khác biệt rất lớn về tốc độ theo đồng hồ, chúng ta lại có kinh nghiệm dồi dào với trí tuệ tập thể, là những hệ thống hình thành từ số lượng khác nhau các cấu phần cấp độ con người, cùng làm việc với nhau với các mức độ hiệu năng khác nhau. Các công ty, các nhóm làm việc, các mạng tán gẫu, các nhóm vận động, các quốc gia, thậm chí toàn thể loài người, nếu áp dụng một cách nhìn trừu tượng hóa ở mức độ nào đó, đều có thể được xem như các “hệ thống”

được xác định một cách lỏng lẻo, có khả năng giải quyết nhiều loại vấn đề trí tuệ. Từ kinh nghiệm, chúng ta phần nào hiểu được việc những tác vụ khác nhau có thể bị chinh phục dễ dàng như thế nào bằng những nỗ lực của các tổ chức có cơ cấu và quy mô đa dạng.

Trí tuệ tập thể vượt trội trong việc giải quyết các vấn đề đã được chia nhỏ thành từng phần sao cho giải pháp cho những vấn đề nhỏ này có thể được thực hiện song song và được kiểm định độc lập. Những tác vụ như chế tạo tàu vũ trụ con thoi hoặc điều hành một nhãn hiệu hamburger nhượng quyền kinh doanh cung cấp vô số cơ hội cho việc phân công lao động: những kỹ sư khác nhau làm việc với những cấu phần khác nhau của con tàu vũ trụ; những nhân viên khác nhau điều hành những nhà hàng khác nhau. Trong giới hàn lâm, sự phân công khắt khe giữa các nhà nghiên cứu, sinh viên, tạp chí, quỹ tài trợ và phần thưởng thành những bộ môn riêng biệt – mặc dù không phù hợp với loại công việc được nói đến trong cuốn sách này – có thể (trong khuôn khổ một lối tư duy “dĩ hòa vi quý”) được xem là sự thích ứng cần thiết đối với tính thực tiễn của việc cho phép một số lượng lớn các cá nhân và nhóm với động cơ đa dạng đóng góp vào sự tăng trưởng kiến thức của loài người trong khi làm việc tương đối độc lập, tự “cày cấy” trên thửa ruộng của mình.

Trí tuệ tập thể của một hệ thống có thể được mở rộng bằng cách tăng số lượng hoặc chất lượng của các trí tuệ thành phần, hoặc cải thiện cách tổ chức của các trí tuệ thành phần đó.⁸ Để có được siêu trí tuệ tập thể từ bất cứ dạng trí tuệ tập thể hiện tại nào, cần phải có một mức độ tăng cường rất lớn. Hệ thống đích sẽ phải có khả năng sở hữu hiệu năng vượt xa bất cứ trí tuệ tập thể hiện tại nào, hoặc vượt xa hệ thống nhận thức khác trên nhiều lĩnh vực rất phổ thông. Bản

thân một định dạng hội thảo mới cho phép các học giả trao đổi thông tin hiệu quả hơn, hay một thuật toán lọc thông tin cộng tác có thể dự báo tốt hơn về đánh giá của người dùng đối với các cuốn sách và bộ phim, rõ ràng không thể tạo ra bất cứ thứ gì tiệm cận siêu trí tuệ tập thể. Việc dân số thế giới tăng 50% hoặc một cải thiện về phương pháp sư phạm cho phép sinh viên đi học một ngày bốn thay vì sáu giờ cũng không thể làm thế. Chúng ta cần phải có một sự tăng trưởng đồ sộ hơn rất nhiều trong năng lực nhận thức tập thể của nhân loại để đáp ứng được tiêu chí của siêu trí tuệ tập thể.

Lưu ý rằng giá trị ngưỡng đối với siêu trí tuệ tập thể có liên hệ với các mức độ hiệu năng của thời điểm hiện tại – có nghĩa là những năm đầu của thế kỷ 21. Trong thời tiền sử của loài người và một lần nữa, trong suốt quá trình lịch sử, trí tuệ tập thể của nhân loại đã và đang tăng trưởng theo phép nhân với thừa số vô cùng lớn. Chẳng hạn, dân số thế giới đã tăng ít nhất một ngàn lần kể từ thế Pleistocene.⁹ Chỉ trên cơ sở này, các mức độ hiện tại của trí tuệ tập thể của loài người có thể được coi như đang tiếp cận được siêu trí tuệ so với vạch mốc là thế Pleistocene. Một số cải thiện về công nghệ truyền thông – đặc biệt là tiếng nói, nhưng cũng có thể là các thành phố, chữ viết và in ấn – có thể được coi như có đem lại những kích thích siêu mạnh một cách riêng lẻ hoặc tổ hợp. Điều này được hiểu là nếu một đổi mới sáng tạo khác có tác động tương đương lên năng lực giải quyết vấn đề của trí tuệ tập thể của chúng ta xuất hiện, thì nó sẽ tạo ra siêu trí tuệ tập thể.¹⁰

Một số bạn đọc nhất định lúc này có thể sẽ muốn ngắt lời chúng tôi và nói rằng xã hội xem ra không thông minh đến như vậy. Có lẽ một số quyết định chính trị không được chào đón nào đó đã được

đưa ra tại quê nhà của các bạn, và sự kém thông thái rõ ràng của quyết định đó giờ đây hiển hiện trong tâm trí bạn đọc như bằng chứng của sự bất lực về tinh thần của kỷ nguyên hiện đại. Và đó chẳng phải là nhân loại hiện giờ đang thần tượng hóa việc tiêu dùng vật chất, làm cạn kiệt tài nguyên thiên nhiên, gây ô nhiễm môi trường hay hủy hoại sự đa dạng sinh học, đồng thời thất bại trong việc giải quyết sự bất công hiện đang gào thét khắp nơi, bỏ qua các giá trị nhân bản và tinh thần to lớn hay sao? Tuy nhiên, đặt sang một bên sự so sánh giữa những thiếu sót của thời kỳ hiện đại với những sai lầm không lấy gì làm nhỏ bé của những thời kỳ trước, không có gì trong định nghĩa của chúng ta về siêu trí tuệ tập thể ám chỉ rằng một xã hội với trí tuệ tập thể lớn hơn sẽ tốt hơn. Thậm chí, định nghĩa này còn không ám chỉ rằng xã hội có trí tuệ tập thể cao hơn thì thông thái hơn. Chúng ta có thể nghĩ về sự thông thái như khả năng hiểu và làm được gần như chuẩn xác những điều quan trọng. Vậy thì khi đó, một tổ chức hình thành từ một lực lượng rất lớn lao động trí thức được điều phối rất hiệu quả, những người có thể cùng nhau giải quyết các vấn đề trí tuệ xuyên suốt nhiều lĩnh vực phổ thông là hoàn toàn có thể hình dung ra được. Tổ chức này, như chúng tôi giả định, có thể vận hành được hầu hết các loại hình doanh nghiệp, sáng chế ra hầu hết các loại công nghệ và tối ưu hóa hầu hết các quy trình. Ngay cả vậy, nó vẫn có thể hoàn toàn sai lầm trong một số ít các vấn đề lớn và bao quát – ví dụ, nó có thể không đủ cẩn trọng khi xử lý những rủi ro mang tính sống còn – và kết quả là sẽ chạy theo một cuộc tăng trưởng nước rút bùng nổ ngắn ngủi, sau đó kết thúc không mấy vinh quang trong sự sụp đổ hoàn toàn. Một tổ chức như vậy có thể có trí tuệ tập thể ở mức độ rất cao; nếu đủ cao thì sẽ là một siêu

trí tuệ tập thể. Chúng ta cần phải tự cưỡng lại ham muốn của mình trong việc gói gọn mọi thuộc tính tiêu chuẩn muốn có thành một khái niệm vô định hình khổng lồ của hoạt động tinh thần, như thể ta không bao giờ tìm được một tính trạng đáng quý của mình mà không có sự hiện diện tương đương của các tính trạng khác. Thay vào đó, chúng ta phải công nhận rằng có thể tồn tại những hệ thống xử lý thông tin mạnh mẽ về mặt công cụ – những hệ thống thông minh – nhưng không có bản chất tốt đẹp hay thông thái một cách đáng tin cậy. Nhưng chúng ta sẽ quay lại chủ đề này trong Chương 7.

Siêu trí tuệ tập thể có thể được tích hợp lỏng lẻo hay chặt chẽ. Để minh họa một trường hợp siêu trí tuệ tích hợp lỏng lẻo, hãy tưởng tượng về một hành tinh tên là MegaEarth (Siêu Trái đất) với cùng công nghệ truyền thông và điều phối mà chúng ta đang có trên Trái đất thực nhưng lại có dân số lớn hơn một triệu lần. Với dân số lớn như vậy, tổng lực lượng lao động trí thức trên MegaEarth cũng lớn hơn tương ứng so với hành tinh của chúng ta. Giả sử một thiên tài khoa học tầm cỡ Newton hay Einstein xuất hiện một lần trên 10 tỷ người, khi đó trên MegaEarth sẽ có 700.000 thiên tài như vậy đang sống, cùng với một số lượng vô cùng lớn các tài năng chỉ kém hơn đôi chút. Những ý tưởng và công nghệ mới có thể sẽ được phát triển với tốc độ điên cuồng, và nền văn minh toàn cầu trên MegaEarth có thể tạo thành một siêu trí tuệ tập thể tích hợp lỏng lẻo.¹¹

Nếu chúng ta tăng dần mức độ thống nhất của một trí tuệ tập thể, cuối cùng nó có thể trở thành một trí tuệ hợp nhất – một “trí óc” lớn duy nhất đối lập với một tập hợp đơn thuần các trí óc nhỏ hơn của con người tương tác lỏng lẻo với nhau.¹² Những cư dân MegaEarth có thể bước đi theo hướng này bằng cách cải thiện các công nghệ

truyền thông và điều phối, và phát triển những phương cách tốt hơn để nhiều cá nhân cùng giải quyết các vấn đề trí tuệ khó khăn. Vì thế, một siêu trí tuệ tập thể sau khi có được tích hợp đủ có thể trở thành một “siêu trí tuệ chất lượng” (quality superintelligence).

Siêu trí tuệ chất lượng

Chúng ta có thể phân ra một loại hình thứ ba của siêu trí tuệ.

Siêu trí tuệ chất lượng: Một hệ thống có tốc độ ít nhất tương đương trí óc con người và thông minh hơn rất nhiều về chất.

Giống với trí tuệ tập thể, chất lượng trí tuệ cũng là một khái niệm khá mơ hồ; và trong trường hợp này, khó khăn còn lớn hơn do chúng ta còn thiếu kinh nghiệm với bất cứ biến thể nào của chất lượng trí tuệ cao hơn mức trần phân phối trí tuệ hiện tại của con người. Tuy nhiên chúng ta có thể có được một số cảm nhận về ý niệm này bằng cách nghiên cứu một số trường hợp có liên quan.

Trước hết, chúng ta có thể mở rộng phạm vi các điểm tham chiếu của mình bằng cách nghiên cứu các loài động vật khác con người có trí tuệ chất lượng thấp. (Điều này không đồng nghĩa với một nhận xét mang tính phân biệt giống loài. Một con cá ngựa vằn có chất lượng trí tuệ phù hợp hoàn hảo với các nhu cầu sinh học của nó; nhưng quan điểm có liên quan ở đây mang tính “dĩ nhân vi trung” (anthropocentric) nhiều hơn: mối quan tâm của chúng ta liên quan đến hiệu năng trong những tác vụ nhận thức phức tạp liên quan đến con người.) Những loài động vật không phải con người không có thứ ngôn ngữ phức tạp có cấu trúc; không có hoặc chỉ có khả năng sơ đẳng trong việc sử dụng hoặc chế tạo công cụ; bị hạn chế nghiêm trọng về khả năng lập kế hoạch dài hạn; và có năng lực lập luận trừu tượng rất hạn chế. Những hạn chế này cũng không thể được giải

thích đầy đủ bởi sự thiếu vắng của trí tuệ tốc độ hoặc trí tuệ tập thể trong tư duy của các loài động vật không phải con người. Về mặt công suất tính toán sơ bộ, não người có lẽ còn kém hơn so với não một số động vật lớn, bao gồm voi và cá voi. Và cho dù nền văn minh công nghệ phức tạp của loài người có thể không có được nếu đi thiếu lợi thế lớn lao về trí tuệ tập thể, thì không phải toàn bộ các khả năng nhận thức riêng biệt của con người đều phụ thuộc vào trí tuệ tập thể. Nhiều năng lực đã được phát triển trong những nhóm săn bắn-hái lượm nhỏ lẻ và cô lập.¹³ Bên cạnh đó, nhiều năng lực khác lại chưa thể phát triển ở mức cao như con người ở một số loài động vật có tổ chức chặt chẽ, ví dụ như tinh tinh hoặc cá heo được con người huấn luyện đặc biệt, hoặc những con kiến sống trong các xã hội lớn và có tôn ti trật tự nghiêm ngặt. Rõ ràng những thành tựu trí tuệ đáng kể của Homo sapiens có được phần lớn là nhờ vào những tính năng đặc thù của kiến trúc não bộ, những tính năng phụ thuộc vào những ưu đãi di truyền đặc thù mà không loài động vật nào sở hữu. Quan sát này có thể giúp chúng tôi minh họa khái niệm siêu trí tuệ chất lượng: đó là trí tuệ có chất lượng vượt trội hơn so với trí tuệ con người ít nhất là tương đương với cách mà trí tuệ con người vượt trội hơn voi, cá heo hay tinh tinh.

Cách thứ hai để minh họa khái niệm siêu trí tuệ chất lượng là thông qua việc chú ý đến những thiếu hụt nhận thức đặc thù cho từng lĩnh vực có thể ảnh hưởng đến từng cá nhân, cụ thể là những thiếu hụt dẫn đến chứng mất trí nhớ tổng thể hoặc những tình trạng khác liên quan đến sự tàn phá quy mô lớn trên những nguồn tài nguyên điện toán thần kinh của não bộ. Hãy xem xét những cá nhân mắc các chứng rối loạn phổ tự kỷ, những người thiếu hụt nhận thức

xã hội nghiêm trọng trong khi vẫn hoạt động tốt ở các lĩnh vực nhận thức khác; hoặc những cá nhân không có khả năng âm nhạc bẩm sinh, những người không thể ngân nga hay nhận ra những giai điệu đơn giản nhưng lại hoàn toàn bình thường trong hầu hết các lĩnh vực khác. Nhiều ví dụ tương tự có thể được bổ sung từ tài liệu thần kinh học, trong đó có vô số điển cứu về bệnh nhân mắc các chứng thiếu hụt năng lực phạm vi hẹp do những bất thường về di truyền hay tổn thương não bộ. Những ví dụ này cho thấy người trưởng thành bình thường có những tài năng nhận thức (cognitive talent) đáng kể không chỉ là một chức năng có được từ việc sở hữu đủ định lượng năng lực xử lý thần kinh tổng quát hay thậm chí là đủ định lượng trí tuệ tổng quát, mà còn cần một hệ thống mạch thần kinh biệt hóa. Quan sát này gợi ý ra ý tưởng về những tài năng nhận thức khả thể nhưng chưa được hiện thực hóa mà không người nào trong thực tế có được, mặc dù các hệ thống trí tuệ khác – các hệ thống không có công suất tính toán cao hơn não người – có thể nâng cao đáng kể năng lực trong việc thực hiện các tác vụ đa dạng phù hợp về mặt chiến lược một khi đã sở hữu chúng.

Theo đó, nhờ nghiên cứu các loài động vật không phải con người và các cá thể người thiếu hụt khả năng nhận thức trong các lĩnh vực đặc thù, chúng ta có thể hình thành ý niệm nào đó về các mức chất lượng khác nhau của trí tuệ và những khác biệt thực sự do chúng tạo ra. Nếu Homo sapiens thiếu đi (chẳng hạn như) những module nhận thức tạo ra những dạng biểu diễn ngôn ngữ phức tạp, thì họ rất có thể cũng chỉ là một loài khỉ khác sống hài hòa với tự nhiên. Ngược lại, nếu có được một bộ module mới nào đó tạo ra lợi thế so

sánh được với việc có khả năng hình thành những dạng biểu diễn ngôn ngữ phức tạp thì chúng ta sẽ có thể trở thành siêu trí tuệ.

Tầm tiếp cận trực tiếp và gián tiếp

Theo thời gian, siêu trí tuệ thuộc bất kỳ hình thái nào ở trên có thể phát triển được công nghệ cần thiết để tạo ra bất cứ hình thái nào khác. Tầm tiếp cận gián tiếp của ba hình thái siêu trí tuệ này như vậy là tương đương. Theo đó, tầm tiếp cận gián tiếp của trí tuệ con người hiện tại cũng nằm trong một lớp tương đương, với giả định rằng cuối cùng chúng ta sẽ có thể tạo ra một hình thái siêu trí tuệ. Nhưng có một chiều hướng mà ở đó ba hình thái siêu trí tuệ này gần gũi với nhau hơn rất nhiều: bất cứ hình thái nào trong số này cũng có thể tạo ra những hình thái siêu trí tuệ khác nhanh hơn so với việc chúng ta có thể tạo ra bất cứ hình thái siêu trí tuệ nào từ điểm xuất phát hiện tại của mình.

Tầm tiếp cận trực tiếp của ba hình thái siêu trí tuệ khác nhau này khó so sánh hơn và có thể không có trật tự xác định. Những năng lực tương ứng của chúng phụ thuộc vào mức độ chúng khởi tạo các lợi thế tương ứng – một siêu trí tuệ tốc độ nhanh đến mức nào, một siêu trí tuệ chất lượng vượt trội về chất lượng đến mức nào, và các vấn đề khác. Chúng ta chỉ có thể nói rằng khi các điều kiện khác không đổi, siêu trí tuệ tốc độ vượt trội trong những tác vụ cần thực thi nhanh chóng một chuỗi dài các bước phải làm tuần tự, trong khi siêu trí tuệ tập thể lại vượt lên trong những tác vụ chấp nhận việc phân tách thành các tác vụ nhỏ song song và những tác vụ đòi hỏi tổ hợp nhiều quan điểm và nhóm kỹ năng. Trong một chiều hướng mơ hồ hơn, siêu trí tuệ chất lượng có thể là hình thái cao nhất, vì nó có thể nắm bắt và giải quyết các vấn đề mà với toàn bộ các mục đích thực tiễn,

nằm ngoài tầm tiếp cận trực tiếp của siêu trí tuệ tốc độ và siêu trí tuệ tập thể.¹⁴

Trong một số lĩnh vực, số lượng là một sự thay thế méo mó cho chất lượng. Một thiên tài cô độc làm việc trong một phòng ngủ bọc gỗ xộp cách âm có thể viết bộ tiểu thuyết *In Search of Lost Time* (Đi tìm thời gian đã mất). Liệu một kiệt tác tương tự có thể được tạo ra bằng cách sử dụng nguyên một tòa nhà văn phòng chứa đầy những tay viết mướn?¹⁵ Thậm chí trong phạm vi biến thể hiện tại của con người, chúng ta thấy rằng một số chức năng có được lợi ích lớn từ công sức của một người chủ trì xuất sắc thay vì nỗ lực chung của vô số người trung bình. Nếu mở rộng tầm hiểu biết của mình để đưa vào đó cả những tư duy siêu trí tuệ, chúng ta sẽ phải đối mặt với khả năng có những vấn đề trí tuệ chỉ có thể giải quyết được bởi siêu trí tuệ và không phù hợp với bất cứ tập thể người không được bổ sung năng lực nào cho dù lớn tới đâu.

Vì vậy, một số vấn đề có thể được giải quyết bằng một siêu trí tuệ chất lượng hay một siêu trí tuệ tốc độ, nhưng một siêu trí tuệ tích hợp lỏng lẻo lại không giải quyết được (trừ khi nó tự tăng cường trí tuệ cho mình trước¹⁶). Chúng ta không thể thấy được rõ ràng tất cả những vấn đề này, nhưng có thể mô tả chúng bằng những thuật ngữ chung.¹⁷

Chúng thường là những vấn đề liên quan đến nhiều mối quan hệ tương thuộc phức tạp không cho phép áp dụng các bước giải quyết độc lập và có thể kiểm định, mà vì thế không thể giải quyết từng phần, và có thể đòi hỏi những hiểu biết hoặc khung biểu diễn thông tin mới về chất có thể quá sâu sắc hay phức tạp để những phiên bản hiện tại của những “người trần mắt thịt” có thể khám phá hoặc sử

dụng hiệu quả. Một số kiểu sáng tạo nghệ thuật và nhận thức chiến lược cũng nằm trong thể loại này, và một số loại hình đột phá khoa học có thể cũng như vậy. Người ta cũng có thể cho rằng sự chậm chạp và lung lay của tiến bộ nhân loại trong nhiều “vấn đề vĩnh cửu” của triết học phát sinh vì sự không thích hợp của não người đối với việc nghiên cứu triết học. Theo quan điểm này, các nhà triết học nổi tiếng nhất của chúng ta cũng giống như những chú chó đi bằng chân sau – rốt cuộc chỉ mới đạt được cấp độ ngưỡng cửa của hiệu năng cần thiết để được tham gia vào các hoạt động này mà thôi.¹⁸

Nguồn tạo lợi thế cho trí tuệ số

Những thay đổi nhỏ về thể tích và kết nối của não bộ có thể có các hệ quả quan trọng, như chúng ta thấy khi so sánh những thành tựu trí tuệ và công nghệ của người và vượn. Những thay đổi lớn hơn nhiều về các nguồn tài nguyên và kiến trúc tính toán mà trí tuệ máy cho phép xảy ra sẽ có thể có những hệ quả thậm chí còn to lớn hơn. Sẽ rất khó, mà có lẽ là không thể để chúng ta hình thành được một cảm nhận trực giác về năng lực của siêu trí tuệ; nhưng ít nhất chúng ta có thể có được gợi ý về không gian của các khả năng bằng cách nhìn vào một số lợi thế mở ra cho các trí óc kỹ thuật số. Nhưng lợi thế về phần cứng là dễ thấy nhất:

- Tốc độ của các phần tử tính toán. Các neuron sinh học vận hành với tốc độ tối đa khoảng 200 Hz, chậm hơn vi xử lý hiện đại (~ 2 GHz) tới 7 cấp khuếch đại.¹⁹

Hệ quả là não người buộc phải trông cậy vào việc vận hành song song quy mô lớn và không thể nhanh chóng thực hiện bất cứ tính toán nào đòi hỏi số lượng lớn các phép tính tuần tự.²⁰ (Mọi thứ được não bộ làm trong khoảng thời gian dưới một giây không thể sử dụng

nhiều hơn 100 phép tính tuần tự, mà có lẽ chỉ khoảng vài chục). Mặc dù vậy, nhiều thuật toán quan trọng nhất trong thực tiễn lập trình và khoa học máy tính lại không thể song song hóa dễ dàng. Nhiều tác vụ nhận thức có thể được thực hiện hiệu quả hơn nhiều nếu sự hỗ trợ tự nhiên của não bộ đối với các thuật toán làm khớp mẫu hình có thể song song hóa được hỗ trợ bởi, và tích hợp với, sự hỗ trợ cho quá trình xử lý tuần tự tốc độ cao.

- Tốc độ giao tiếp nội bộ. Dây thần kinh truyền hiệu lệnh hành động với tốc độ 120 m/s hoặc chậm hơn, trong khi đó các lõi xử lý điện tử có thể liên lạc với nhau qua đường quang học với tốc độ ánh sáng (300.000.000 m/s).²¹ Sự chậm trễ của tín hiệu thần kinh hạn chế kích thước của bộ não sinh học trong khi nó vận hành như một đơn vị xử lý duy nhất. Ví dụ, để đạt được độ trễ khứu hồi 10 ms giữa hai phần tử bất kỳ trong một hệ thống, các bộ não sinh học phải có thể tích nhỏ hơn 0,11 m³. Một hệ thống điện tử, mặt khác, có thể có thể tích lên tới 6,1 x 10¹⁷ m³, gần bằng kích thước của một hành tinh lùn: lớn hơn 18 cấp khuếch đại.

- Số lượng phần tử tính toán. Não người có chưa tới 100 triệu neuron,²³ và có kích thước lớn hơn não tinh tinh ba lần rưỡi (nhưng chỉ bằng một phần năm cá nhà táng).²⁴ Số neuron trong một sinh vật chủ yếu bị giới hạn bởi thể tích hộp sọ và những hạn chế về chuyển hóa, nhưng các yếu tố khác cũng khá quan trọng đối với những bộ não lớn (như làm mát, thời gian phát triển và độ trễ dẫn truyền tín hiệu – xem ý trước). Ngược lại, phần cứng máy tính có khả năng mở rộng vô biên tới những giới hạn vật lý rất cao.²⁵ Các siêu máy tính có thể lớn như một nhà kho hoặc hơn nữa, với dung lượng bổ sung từ xa thông qua cáp tốc độ cao.²⁶

- Dung lượng lưu trữ. Bộ nhớ làm việc của con người không thể lưu nhiều hơn bốn hoặc năm khối thông tin ở bất kỳ thời điểm nào.²⁷ Trong khi việc so sánh trực tiếp kích thước bộ nhớ làm việc của con người với RAM của một máy tính kỹ thuật số có thể là khập khiễng, rõ ràng lợi thế phần cứng của trí tuệ kỹ thuật số có thể giúp chúng có bộ nhớ làm việc lớn hơn. Điều đó cho phép các trí tuệ này nắm bắt bằng trực giác những mối quan hệ phức tạp mà con người chỉ có thể xử lý một cách vụng về thông qua những phép tính toán nặng nhọc.²⁸ Trí nhớ dài hạn của con người cũng bị hạn chế, mặc dù chưa rõ liệu chúng ta có sử dụng được hết dung lượng của nó trong phần đời thông thường hay không – tốc độ tích lũy thông tin của chúng ta quá chậm (theo một đánh giá, não người lưu trữ khoảng một tỷ bit – nhỏ hơn một, hai cấp khuếch đại so với một chiếc điện thoại thông minh cấp thấp.²⁹) Cả lượng thông tin được lưu trữ và tốc độ truy nhập thông tin trong một bộ não máy do đó cao hơn một bộ não sinh học rất nhiều.

- Độ tin cậy, tuổi thọ, các cảm biến,... Các trí tuệ máy có thể có những lợi thế khác về phần cứng. Ví dụ, các neuron sinh học có độ tin cậy kém hơn các bóng bán dẫn.³⁰ Vì việc tính toán có độ nhiễu cao cần đến những phương thức mã hóa dư thừa sử dụng nhiều phần tử để mã hóa một bit thông tin, nên một bộ não kỹ thuật số có thể sẽ tăng được phần nào hiệu năng từ việc sử dụng những phần tử tính toán đáng tin cậy và có độ chính xác cao. Các bộ não trở nên mệt mỏi sau vài giờ làm việc và bắt đầu sa sút sau một vài thập kỷ tính bằng thời gian chủ quan, trong khi các bộ vi xử lý không bị ảnh hưởng bởi những hạn chế này. Luồng dữ liệu chạy trong một trí tuệ máy có thể được tăng lên bằng cách bổ sung hàng triệu bộ cảm biến.

Tùy theo công nghệ được sử dụng, phần mềm của một chiếc máy có thể được cấu hình nhằm tối ưu hóa theo những yêu cầu tác vụ thay đổi, trong khi đó phần lớn kiến trúc não bộ đã bị cố định hoặc chỉ có thể thay đổi chậm chạp (mặc dù các chi tiết của mạng lưới kết nối synapse thay đổi sau những thang thời gian ngắn hơn, theo ngày chẳng hạn.³¹⁾

Hiện nay, công suất tính toán của bộ não sinh học vẫn còn cao hơn máy tính số, mặc dù các siêu máy tính hàng đầu đã đạt đến những cấp độ hiệu năng nằm trong phạm vi giá trị ước tính hợp lý về năng lực xử lý của bộ não.³² Tuy nhiên, phần cứng đang được cải thiện nhanh chóng, và những giới hạn cuối cùng của hiệu năng phần cứng là cao hơn rất nhiều so với các chất nền tính toán sinh học.

Trí não số cũng được lợi từ những tiến bộ lớn về phần mềm:

- Tính dễ sửa đổi. Thí nghiệm với các biến thể tham số trong phần mềm dễ hơn nhiều so với phần ướt thần kinh. Ví dụ, với một bản giả lập hoàn chỉnh não bộ, ta có thể dễ dàng thí nghiệm xem chuyện gì xảy ra nếu bổ sung thêm neuron trong một vùng vỏ não cụ thể nào đó hoặc nếu tăng hay giảm khả năng kích thích của chúng. Thực hiện những thí nghiệm này trên bộ não sinh học sống sẽ khó khăn hơn rất nhiều.

- Khả năng sao chép. Với phần mềm, ta có thể nhanh chóng tạo ra rất nhiều bản sao độ trung thực cao để lấp đầy cơ sở phần cứng có sẵn. Các bộ não sinh học, ngược lại, chỉ có thể tái tạo với tốc độ rất chậm; và mỗi bản mới lại khởi đầu từ tình trạng “vô năng”, không nhớ được bất cứ thứ gì mà các bậc cha mẹ đã học trong cuộc sống.

- Phối hợp mục tiêu. Các tập thể người luôn chứa đầy sự kém hiệu quả phát sinh từ việc gần như không thể đạt được sự đồng

thuận về mục tiêu giữa các thành viên của một nhóm lớn – ít nhất cho tới khi có thể tạo ra sự “dễ bảo” trên quy mô lớn bằng thuốc hay chọn lọc di truyền. Một “băng đảng sao chép” (copy clan, một nhóm các chương trình đồng nhất hay gần như đồng nhất có chung một mục tiêu) sẽ tránh được những vấn đề phối hợp như vậy.

- Dùng chung bộ nhớ. Bộ não sinh học cần các khoảng thời gian đào tạo và cố vấn kéo dài, còn trí óc kỹ thuật số có thể lấy được những ký ức và kỹ năng mới bằng cách trao đổi các file dữ liệu. Tổ hợp hàng tỷ bản sao của một chương trình AI có thể định kỳ đồng bộ cơ sở dữ liệu của mình sao cho tất cả các bản sao đều biết mọi thứ mà bất cứ bản nào khác học được trong khung giờ trước đó. (Việc truyền bộ nhớ trực tiếp cần có các định dạng biểu diễn được chuẩn hóa. Việc hoán đổi dễ dàng nội dung nhận thức ở cấp độ cao do đó có thể không thực hiện được giữa các cặp trí tuệ máy bất kỳ. Cụ thể, việc này có thể không thực hiện được giữa các bản giả lập hoàn chỉnh não bộ thế hệ đầu tiên.)

- Những module, thể thức và thuật toán mới. Tiếp nhận thông tin thị giác với chúng ta có vẻ dễ dàng và không cần nỗ lực, rất khác so với việc giải các bài toán hình học trong sách giáo khoa – mặc dù cần đến một khối lượng tính toán khổng lồ để tái tạo một dạng biểu diễn ba chiều của thế giới chứa đầy những vật thể dễ nhận biết từ những mẫu hình mô phỏng hai chiều trên võng mạc. Việc này có vẻ dễ dàng vì chúng ta dành riêng những máy móc thần kinh cấp độ thấp cho việc xử lý thông tin hình ảnh. Việc xử lý cấp thấp này diễn ra vô thức và tự động, không tiêu thụ năng lượng thần kinh hay sự chú ý có ý thức của chúng ta. Cảm nhận âm nhạc, sử dụng ngôn ngữ, nhận thức xã hội và những hình thái xử lý thông tin “tự nhiên” khác của

con người xem ra cũng được hỗ trợ tương tự bởi những module tính toán thần kinh chuyên biệt. Một trí óc nhân tạo được hỗ trợ chuyên biệt như vậy cho những lĩnh vực nhận thức khác đang trở nên quan trọng trong thế giới hiện đại – như kỹ thuật, lập trình máy tính và chiến lược kinh doanh – sẽ có lợi thế lớn so với những trí óc như của chúng ta, thứ chủ yếu trông cậy vào nhận thức phổ quát và công kênh để suy nghĩ. Những thuật toán mới có thể được phát triển để tận dụng khả năng đặc biệt của phần cứng kỹ thuật số, chẳng hạn như hỗ trợ xử lý tuần tự tốc độ cao.

Những lợi thế cuối cùng có thể có được của trí tuệ máy kết hợp cả phần cứng và phần mềm là vô cùng to lớn.³³ Nhưng những lợi thế tiềm năng này có thể được hiện thực hóa nhanh đến mức nào? Đó chính là câu hỏi tiếp theo mà chúng ta cần giải đáp.

Chương 4

Động lực của sự bùng nổ trí tuệ

Một khi máy móc có được dạng thức nào đó tương đương con người về khả năng lập luận tổng thể, liệu còn bao nhiêu lâu nữa trước khi chúng có được siêu trí tuệ toàn diện? Liệu đó là một quá trình chuyển đổi chậm, dần dần và kéo dài hay sẽ bất ngờ bùng nổ? Chương này sẽ phân tích động lực của quá trình chuyển đổi sang siêu trí tuệ dưới dạng một hàm của công suất tối ưu hóa và sức chống đối của hệ thống. Chúng tôi xem xét những gì mình biết hoặc có thể phỏng đoán một cách hợp lý về hành vi của hai yếu tố này trong vùng lân cận của trí tuệ tổng quát cấp độ con người.

Thời điểm và tốc độ của quá trình “cất cánh”

Xét rằng máy móc cuối cùng sẽ vượt trội hơn sinh học rất nhiều về trí tuệ tổng quát, nhưng nhận thức máy hiện nay lại hạn hẹp hơn nhiều so với con người, ta sẽ phải băn khoăn xem khi nào thì sự “soán ngôi” sẽ xảy ra. Câu hỏi chúng tôi đặt ra ở đây cần được phân biệt rõ ràng với câu hỏi đã nghiên cứu trong Chương 1 về khoảng cách của chúng ta hiện nay tới việc phát triển được một cỗ máy có trí tuệ tổng quát cấp độ con người. Thay vào đó, câu hỏi ở đây là: nếu, và khi cỗ máy đó được phát triển, sẽ tiếp tục mất bao lâu tới khi nó hoàn toàn trở thành siêu trí tuệ? Hãy lưu ý rằng một người có thể nghĩ sẽ phải mất khá nhiều thời gian trước khi máy móc đạt đến

đường cơ sở con người, hoặc có thể tỏ ra “bất khả tri” về khoảng thời gian để điều đó xảy đến, nhưng vẫn có quan điểm mạnh mẽ rằng một khi chuyện đó xảy ra, một siêu trí tuệ mạnh mẽ sẽ xuất hiện rất nhanh chóng.

Suy nghĩ về những vấn đề này bằng sơ đồ có thể hữu ích, tuy cũng sẽ tạm thời bỏ qua một số nhận định và chi tiết làm phức tạp hóa vấn đề. Vậy hãy xem xét một đồ thị biểu diễn năng lực trí tuệ của hệ thống trí tuệ máy tiên tiến nhất như một hàm thời gian (Hình 7).

Một đường thẳng ngang gọi là “đường cơ sở con người” biểu diễn những năng lực trí tuệ hiệu dụng của một người trưởng thành đại diện có thể tiếp cận các nguồn thông tin và công cụ trợ giúp công nghệ hiện có ở các nước phát triển. Hiện tại, hệ thống AI tiên tiến nhất vẫn đang ở mức thấp hơn nhiều so với đường cơ sở con người, dựa trên bất cứ tiêu chí hợp lý nào của năng lực trí tuệ tổng quát. Tại một điểm nào đó trong tương lai, một chiếc máy có thể đạt được trình độ gần như đồng mức với đường cơ sở con người này. (Chúng ta coi đường cơ sở này là cố định, lấy mốc là năm 2014, ngay cả nếu năng lực của các cá nhân có thể tăng lên trong những năm sau đó): điều đó sẽ đánh dấu điểm khởi đầu của quá trình “cất cánh”. Các năng lực của hệ thống tiếp tục tăng trưởng, và đến một thời điểm nào đó muộn hơn, hệ thống sẽ đạt đến trình độ ngang với năng lực trí tuệ kết hợp của toàn nhân loại (vẫn lấy năm 2014 làm mốc): chúng ta có thể gọi đường này là “đường cơ sở nền văn minh”. Cuối cùng, nếu các năng lực của hệ thống tiếp tục tăng trưởng, nó sẽ đạt được “siêu trí tuệ mạnh” – một cấp độ trí tuệ cao hơn rất nhiều so với năng lực trí tuệ kết hợp của toàn thể loài người hiện tại. Việc đạt được siêu trí tuệ mạnh đánh dấu sự hoàn tất của quá trình cất cánh,

mặc dù sau đó hệ thống có thể tiếp tục tăng trưởng năng lực. Đôi khi trong giai đoạn cắt cánh, hệ thống có thể đi qua một cột mốc gọi là “điểm quá độ” (crossover), một điểm mà khi vượt qua thì sự cải thiện hơn nữa của hệ thống chủ yếu được thực hiện bởi hành động của chính hệ thống thay vì công sức của các đối tượng khác.¹ (Sự tồn tại khả thể của điểm chuyển tiếp sẽ trở nên quan trọng trong phần sau của chương này nói về công suất tối ưu hóa và khả năng bùng nổ.)



Hình 7: Dạng đồ thị của quá trình “cắt cánh”. Các câu hỏi sau đây cần phải được khu biệt: “Quá trình cắt cánh có xảy ra không, và nếu có thì khi nào?” và “Nếu và khi quá trình cắt cánh xảy ra, độ dốc đồ thị sẽ ra sao?” Ví dụ, một người có thể cho rằng sẽ còn rất lâu nữa mới xảy ra một lần cắt cánh, nhưng khi đó nó sẽ tiến triển rất nhanh chóng. Một câu hỏi có liên quan khác (không được minh họa trong hình này) là: “Bộ phận của nền kinh tế thế giới tham gia quá trình cắt cánh này sẽ lớn đến đâu?” Những câu hỏi này có liên quan đến nhau, nhưng vẫn là những vấn đề khác biệt.

Với bức tranh đó trong đầu, chúng ta có thể phân biệt ba lớp kịch bản chuyển đổi – các kịch bản mà trong đó các hệ thống phát triển từ trí tuệ cấp độ con người lên siêu trí tuệ – dựa trên độ dốc của chúng; có nghĩa là chúng biểu diễn quá trình cắt cánh có tốc độ chậm, nhanh hay trung bình.

Chậm

Quá trình cắt cánh chậm là quá trình cắt cánh xảy ra trong một khoảng thời gian dài, có thể hàng thập kỷ hoặc hàng thế kỷ. Những

kịch bản cốt cánh chậm tạo ra các cơ hội tuyệt vời để các quá trình chính trị của loài người có thể thích nghi và phản ứng lại được. Những cách tiếp cận khác nhau có thể được thử nghiệm và kiểm tra lần lượt, những chuyên gia mới có thể được đào tạo và công nhận, và các chiến dịch cấp thấp có thể được huy động bởi các nhóm cảm thấy rằng mình bị yếu thế trước những phát triển đang phô bày trước mắt. Nếu cơ sở hạ tầng bảo mật mới hoặc sự giám sát đại trà của các nhà nghiên cứu AI tỏ ra cần thiết, những hệ thống này sẽ được phát triển và triển khai. Các dân tộc lo sợ một cuộc chạy đua vũ trang AI sẽ có thời gian để tìm cách đàm phán các hiệp ước và thiết kế ra các cơ chế thực thi. Hầu hết việc chuẩn bị được thực hiện trước khi quá trình cốt cánh chậm bắt đầu xảy ra có thể sẽ trở nên lạc hậu khi những giải pháp tốt hơn dần hiển hiện trong buổi bình minh của kỷ nguyên mới.

Nhanh

Quá trình cốt cánh nhanh diễn ra trong một khoảng thời gian ngắn, ví dụ trong vài phút, vài giờ hay vài ngày. Các kịch bản cốt cánh nhanh không cho con người có đủ cơ hội để nghiên cứu. Sẽ không có ai kịp nhận ra điều bất thường trước khi cuộc chơi kết thúc. Trong kịch bản cốt cánh nhanh, số phận của nhân loại về cơ bản phụ thuộc vào những chuẩn bị đã thực hiện trước đó. Ở trạng thái chậm nhất của kịch bản cốt cánh nhanh, một số hành động đơn giản của con người có thể khả thi, giống như mở “chiếc va li nguyên tử”, nhưng bất cứ hành động nào như vậy cũng sẽ chỉ là sơ đẳng hoặc đã được lên kế hoạch và lập trình từ trước.

Trung bình

Quá trình cất cánh trung bình diễn ra trong một khoảng thời gian trung bình, như vài tháng hoặc vài năm. Các kịch bản cất cánh trung bình cho loài người một vài cơ hội phản ứng nhưng không nhiều thời gian để phân tích tình huống, để thử nghiệm những cách tiếp cận khác nhau, hoặc để giải quyết những vấn đề phối hợp phức tạp. Tuy chúng ta không có đủ thời gian để phát triển hoặc triển khai những hệ thống mới (chẳng hạn như các hệ thống chính trị, các chế độ giám sát hoặc các giao thức bảo mật mạng máy tính), nhưng những hệ thống còn lại cũng có thể được áp dụng vào những thách thức mới.

Trong quá trình cất cánh chậm, sẽ có vô khối thời gian để tin tức được lan truyền. Ngược lại, trong một cuộc cất cánh trung bình, có khả năng những diễn biến sẽ được giữ bí mật khi chúng diễn ra. Thông tin có thể bị giới hạn trong một nhóm nhỏ những “người trong cuộc”, chẳng hạn như trong một chương trình nghiên cứu quân sự do chính phủ tài trợ. Những lớp vỏ bọc là các dự án thương mại, những nhóm nghiên cứu nhỏ và “chín tay tin tặc dưới hầm ngầm” có thể cũng được giữ bí mật – mặc dù nếu như viễn cảnh của một sự bùng nổ trí tuệ bị “rơi vào ra đa” của các cơ quan tình báo nhà nước như một ưu tiên về an ninh quốc gia, thì những dự án tư nhân hứa hẹn nhất rất có thể sẽ bị giám sát. Nhà nước sở tại (hoặc một thế lực thống trị ngoại bang nào đó) khi đó có thể có lựa chọn quốc hữu hóa hoặc đóng cửa bất cứ dự án nào có dấu hiệu bắt đầu cất cánh. Quá trình cất cánh nhanh có thể xảy ra nhanh tới mức không còn đủ thời gian để thế giới có thể thoát thân, hoặc ai đó có thể có phản ứng đáng kể. Nhưng người ngoài cuộc có thể can thiệp trước khi quá trình cất cánh xảy ra, nếu họ tin rằng một dự án cụ thể đang tiến gần đến thành công.

Các kịch bản cắt cánh trung bình có thể dẫn đến sự xáo động về địa chính trị, xã hội và kinh tế khi các cá nhân và nhóm tự tranh đấu để có thể thu lợi từ việc chuyển đổi đang diễn ra. Cuộc “nổi dậy” như vậy, nếu như diễn ra, có thể phá hủy những nỗ lực xây dựng một biện pháp phản ứng chặt chẽ; hoặc tạo điều kiện cho những giải pháp cực đoan hơn so với những giải pháp khả thể mà các tình huống ít sóng gió hơn cho phép thực hiện. Ví dụ, trong một kịch bản cắt cánh trung bình mà ở đó những bản giả lập giá rẻ và hiệu quả hoặc những dạng trí óc kỹ thuật số khác dần dần tràn ngập thị trường lao động trong vài năm, ta có thể hình dung ra được những cuộc biểu tình rộng rãi của công nhân bị sa thải gây áp lực buộc chính phủ tăng phúc lợi thất nghiệp hoặc đưa ra một bảo đảm tiền lương đủ sống cho toàn bộ công dân, hoặc đánh những loại thuế đặc biệt hay áp đặt những yêu cầu về mức lương tối thiểu lên các chủ lao động sử dụng công nhân giả lập. Để bất kỳ biện pháp uốn nắn nào có được từ những chính sách này không mang tính tạm thời, sự hỗ trợ đối với chúng cần phải được gắn liền với các cơ cấu quyền lực dài hạn. Những vấn đề tương tự có thể nảy sinh nếu quá trình cắt cánh là chậm thay vì trung bình, nhưng sự mất cân bằng và thay đổi nhanh chóng trong những kịch bản trung bình có thể đem lại cơ hội đặc biệt cho những nhóm nhỏ để có thể có được tầm ảnh hưởng lớn.

Một số độc giả có thể thấy rằng trong số ba kiểu kịch bản này, cắt cánh chậm có xác suất cao nhất, cắt cánh trung bình có xác suất thấp hơn và cắt cánh nhanh là gần như không thể xảy ra. Hẳn sẽ phi thực tế khi giả định rằng thế giới này có thể thay đổi hoàn toàn và loài người sẽ bị hạ bệ khỏi vị thế của loài có tư duy thượng đẳng chỉ

trong vòng một hoặc hai giờ. Sự thay đổi như vậy chưa từng xuất hiện trong lịch sử loài người, và thứ gần nhất với nó – các cuộc Cách mạng Nông nghiệp và Công nghiệp – đã xảy ra trong những thang thời gian dài hơn nhiều (hàng thế kỷ đến hàng thiên niên kỷ với Cách mạng Nông nghiệp và hàng thập kỷ đến hàng thế kỷ cho Cách mạng Công nghiệp). Như vậy, tốc độ cơ sở cho kiểu chuyển đổi gây ra bởi một kịch bản cất cánh nhanh hay trung bình, về mặt tốc độ và độ lớn của thay đổi giả định là 0: nó không có tiền lệ nào nằm ngoài truyền thuyết và tôn giáo.²

Mặc dù vậy, chương này vẫn sẽ đưa ra một số lý do để nghĩ rằng kịch bản chuyển đổi chậm cũng ít có khả năng. Và khi cất cánh xảy ra, khả năng cao sẽ là một sự bùng nổ.

Để bắt đầu phân tích câu hỏi về tốc độ của quá trình cất cánh, chúng ta có thể coi tốc độ gia tăng trí tuệ của một hệ thống như một hàm (tăng đơn điệu) của hai biến: định lượng “công suất tối ưu hóa”, hoặc nỗ lực thiết kế đánh giá theo chất lượng, được áp dụng để gia tăng trí tuệ cho hệ thống, và tính đáp ứng của hệ thống với việc ứng dụng một định lượng năng lực tối ưu hóa cho trước. Chúng ta có thể gọi giá trị nghịch đảo của tính đáp ứng là “sức chống đối” và viết ra biểu thức:

Tốc độ thay đổi trí tuệ = Công suất tối ưu hóa/Sức chống đối

Trước khi tìm ra cụ thể cách thức lượng hóa trí tuệ, nỗ lực thiết kế và sức chống đối thì biểu thức này chỉ mang tính định tính. Nhưng ít nhất, chúng ta cũng có thể quan sát được rằng trí tuệ của hệ thống sẽ tăng nhanh nếu nhiều thao tác thành thạo được áp dụng vào việc gia tăng trí tuệ của nó và trí tuệ của hệ thống không phải là thứ khó gia tăng, hoặc nỗ lực thiết kế là đáng kể và sức chống đối của hệ

thống là thấp, hoặc cả hai. Nếu định lượng được nỗ lực thiết kế áp dụng để cải thiện một hệ thống cụ thể và tốc độ cải thiện mà nỗ lực này tạo ra, chúng ta có thể tính toán được sức chống đối của hệ thống.

Hơn nữa, chúng ta có thể quan sát được định lượng công suất tối ưu hóa dành cho việc cải thiện hiệu năng hệ thống là khác nhau theo thời gian và giữa các hệ thống. Sức chống đối của một hệ thống cũng có thể thay đổi tùy vào việc hệ thống đã được tối ưu hóa đến mức nào. Thông thường, những cải thiện dễ dàng nhất sẽ được thực hiện trước, dẫn đến hiệu suất giảm dần (tăng sức chống đối), vì những “trái” thấp hơn trên “cây” đã dần bị hái hết. Tuy nhiên, có thể có những cải thiện làm cho những cải thiện sau đó trở nên dễ dàng hơn, và tạo ra những “dòng thác” cải thiện. Quá trình giải một bức tranh ghép bắt đầu một cách đơn giản – các góc và cạnh của bức tranh khá dễ tìm. Sau đó, sức chống đối tăng lên và những mảnh ghép tiếp theo sẽ khó lắp hơn. Nhưng khi bức tranh gần hoàn thiện, không gian tìm kiếm sẽ thu nhỏ lại và quá trình ghép tranh lại trở nên dễ dàng.

Vì thế, để có tiến triển trong cuộc tìm kiếm của mình, chúng ta cần phân tích cách thức mà sức chống đối và công suất tối ưu hóa có thể biến thiên trong những khoảng thời gian thiết yếu trong khi cất cánh. Chúng ta sẽ bàn về việc này trong những trang viết sau đây.

Sức chống đối

Chúng ta hãy cùng bắt đầu với sức chống đối. Quan điểm ở đây phụ thuộc vào kiểu hệ thống đang được xem xét. Để đảm bảo đầy đủ, trước hết, chúng tôi xem xét sơ lược những gì sức chống đối có thể gặp phải trên những con đường dẫn tới siêu trí tuệ mà không liên

quan đến trí tuệ máy tiên tiến. Chúng tôi thấy rằng sức chống đối trên những con đường này là tương đối cao. Sau khi hoàn thành việc này, chúng tôi sẽ chuyển sang trường hợp chính, là quá trình cất cánh có liên quan đến trí tuệ máy; và ở đây chúng tôi thấy rằng sức chống đối ở những thời điểm quan trọng xem ra khá thấp.

Các con đường trí tuệ phi máy móc

Việc tăng cường nhận thức thông qua những cải thiện về y tế công cộng và chế độ ăn uống có hiệu suất giảm dần.³ Những ích lợi lớn đến từ việc xóa sổ tình trạng suy dinh dưỡng trầm trọng, và những dạng suy dinh dưỡng nguy hiểm nhất đã bị loại bỏ ở hầu hết tất cả các quốc gia, trừ những quốc gia nghèo nhất. Việc tăng cường chế độ ăn vốn đã phù hợp chỉ làm tăng vòng bụng. Tương tự, giáo dục giờ đây có lẽ cũng gặp phải vấn đề hiệu suất giảm dần. Số lượng cá nhân tài năng trên thế giới không được tiếp cận nền giáo dục có chất lượng vẫn còn đáng kể nhưng đang giảm dần.

Các loại dược phẩm tăng cường có thể đem lại đôi chút tăng ích về nhận thức trong những thập kỷ tiếp theo. Nhưng sau khi những kết quả dễ nhất đã đạt được – có thể là sự tăng trưởng bền vững về năng lực trí óc và khả năng tập trung, cùng với khả năng kiểm soát tốt hơn việc hình thành ký ức dài hạn – những tăng ích tiếp theo có thể khó đạt được. Tuy nhiên, không giống cách tiếp cận y tế công cộng và chế độ ăn uống, cải thiện nhận thức nhờ các loại thuốc thông minh có thể dễ dàng trước khi trở nên khó khăn hơn. Ngành dược phẩm thần kinh vẫn còn thiếu nhiều kiến thức cơ bản có thể cần đến để can thiệp một cách thành thạo vào một bộ não khỏe mạnh. Việc không coi các loại thuốc tăng cường như một lĩnh vực nghiên cứu hợp lệ có thể phần nào là nguyên nhân của sự lạc hậu

này hiện nay. Nếu ngành khoa học thần kinh và dược phẩm tiếp tục phát triển thêm một thời gian nữa mà không tập trung vào việc tăng cường nhận thức, có thể sẽ có được một số kết quả tương đối dễ dàng khi cuối cùng việc phát triển các loại thuốc thông minh trở thành một ưu tiên nghiêm túc.⁴

Việc tăng cường nhận thức di truyền có đồ thị sức chống đối hình chữ U, giống với đồ thị của các loại thuốc thông minh, nhưng với những tăng ích tiềm tàng lớn hơn. Sức chống đối ban đầu là khá cao trong khi phương pháp sẵn có duy nhất là chọn giống bền vững qua nhiều thế hệ, điều mà rõ ràng là rất khó thực hiện trên quy mô toàn cầu. Phương pháp tăng cường di truyền sẽ trở nên dễ dàng hơn khi công nghệ đã được phát triển để kiểm tra và lựa chọn gen hiệu quả với chi phí thấp (đặc biệt là khi phương pháp chọn lọc phôi lặp lại trở nên khả thi trên người). Những kỹ thuật mới này sẽ giúp khai thác vốn gen người đa dạng hiện tại để có các allele tăng cường trí tuệ. Tuy nhiên, khi những allele tốt nhất hiện có đã được đưa vào các gói tăng cường di truyền, những tăng ích tiếp theo sẽ trở nên khó đạt được hơn. Nhu cầu về những cách tiếp cận sáng tạo hơn đối với quá trình sửa đổi di truyền khi đó có thể làm tăng sức chống đối. Tốc độ phát triển của mọi thứ có thể bị giới hạn trên con đường tăng cường di truyền, rõ ràng nhất là các biện pháp can thiệp dòng mầm không thể tránh khỏi bị ảnh hưởng bởi độ trễ trưởng thành: điều này hoàn toàn đi ngược lại với xác suất của quá trình cất cánh nhanh hoặc trung bình.⁵ Việc chọn lọc phôi chỉ có thể thực hiện được bằng thụ tinh ống nghiệm sẽ làm chậm tốc độ ứng dụng, và đây là một nhân tố hạn chế khác.

Sức chống đối trên con đường não-máy tính ban đầu xem ra rất lớn. Ở một trường hợp hiếm gặp mà trong đó việc đặt các thiết bị cấy ghép não và đạt được mức độ tích hợp chức năng cao với vỏ não bằng cách nào đó có thể trở nên dễ dàng, sức chống đối sẽ có thể “lao dốc”. Trong dài hạn, khó khăn của việc đạt được tiến bộ trên con đường này có thể giống với khó khăn liên quan đến cải thiện các bản giả lập hay các AI, vì một phần trí tuệ của hệ thống não-máy tính có thể cuối cùng lại nằm trong phần máy tính.

Sức chống đối trước việc tăng cường hiệu năng của các mạng lưới và tổ chức là tương đối cao. Vô vàn nỗ lực đang được đổ vào quá trình vượt qua sức chống đối này, nhưng chỉ mang lại sự cải thiện về năng lực tổng thể của loài người mỗi năm không quá một, hai phần trăm.⁶

Hơn nữa, những dịch chuyển về môi trường bên trong và bên ngoài đồng nghĩa với việc các tổ chức (ngay cả các tổ chức đã từng có hiệu quả) sẽ nhanh chóng trở nên kém thích ứng với hoàn cảnh mới của mình. Do đó, cần đến nỗ lực cải cách thường xuyên ngay cả khi chỉ để phòng ngừa hư hại. Một thay đổi lớn về tốc độ tăng ích hiệu năng tổ chức trung bình hẳn là rất dễ hình dung, nhưng khó để thấy được làm thế nào để kích bản thậm chí là cực đoan nhất thuộc loại này có thể tạo ra bất cứ thứ gì nhanh hơn một quá trình cắt cánh chậm, vì các tổ chức do con người vận hành đều bị giới hạn trong thang thời gian của con người. Internet tiếp tục là một biên giới đầy phấn khích với nhiều cơ hội tăng cường trí tuệ tập thể, với sức chống đối tại thời điểm này có vẻ như nằm trong phổ trung bình – tốc độ tiến bộ là tương đối nhanh nhưng lại tiêu tốn quá nhiều nỗ lực. Tốc

độ tiến bộ này được kỳ vọng sẽ gia tăng khi những “trái thấp trên cây” (ví dụ các động cơ tìm kiếm và email) đã được hái hết.

Các con đường giả lập và AI

Rất khó ước tính được những khó khăn của việc tiến tới giả lập hoàn chỉnh não bộ. Mặc dù vậy, chúng ta có thể chỉ ra một cột mốc tương lai cụ thể: giả lập thành công một bộ não côn trùng. Cột mốc này nằm trên một cao điểm, và việc chinh phục nó giúp quan sát được nhiều địa hình phía trước, cho phép chúng ta dự đoán tương đối chính xác về sức chống đối trước quá trình mở rộng phạm vi của công nghệ này lên mức giả lập hoàn chỉnh não bộ con người. (Một bản giả lập thành công bộ não của động vật có vú nhỏ, như chuột, có thể tạo ra một điểm quan sát tốt hơn, cho phép đánh giá được khoảng cách còn lại tới việc giả lập hoàn chỉnh não bộ con người với độ chính xác cao.) Con đường hướng tới AI, ngược lại, có thể không có những cột mốc hay những điểm quan sát sớm rõ ràng như vậy. Công cuộc chinh phục AI có thể bị lạc lối hoàn toàn trong rừng rậm, cho tới khi một đột phá bất ngờ làm hé lộ vạch đích trong một vùng trống chỉ cách vài bước chân gần gũi.

Hãy nhớ lại sự khác biệt giữa hai câu hỏi sau đây: Đạt được năng lực tư duy gần bằng mức độ của con người khó đến đâu? Và từ đó đến được các cấp độ siêu nhân khó đến mức nào? Câu hỏi thứ nhất chủ yếu liên quan đến dự báo thời điểm bắt đầu cất cánh, còn câu hỏi thứ hai chính là chìa khóa để đánh giá dạng đồ thị của quá trình cất cánh, và đó chính là mục tiêu của chúng ta ở đây. Và tuy rất muốn giả định rằng bước đi từ cấp độ con người lên mức độ siêu nhân phải là bước đi khó khăn hơn – bước đi này, tựu chung lại, lại diễn ra “trên cao độ lớn hơn” mà ở đó, một hệ thống vốn đã khá hiệu

quả vẫn cần bổ sung năng lực – nhưng đó có thể là một giả định rất không an toàn. Sức chống đối hoàn toàn có thể suy giảm khi máy móc đã đạt được cấp độ tương đương con người.

Hãy xem xét bản giả lập hoàn chỉnh não bộ đầu tiên. Những khó khăn liên quan đến tạo ra con người giả lập đầu tiên là hoàn toàn khác với những khó khăn liên quan đến việc mở rộng một bản giả lập sẵn có. Tạo ra mô phỏng đầu tiên liên quan đến những thách thức công nghệ khổng lồ, cụ thể là về mặt phát triển những năng lực quét và diễn dịch hình ảnh cần thiết. Bước đi này cũng có thể đòi hỏi một lượng vốn vật lý đáng kể – một khu vực tập trung máy móc có quy mô công nghiệp với hàng trăm máy quét công suất cao không phải là một chuyện hoang đường. Ngược lại, việc tăng cường chất lượng của một bản giả lập sẵn có lại liên quan đến việc điều chỉnh các thuật toán và cấu trúc dữ liệu. Về bản chất, đây là một vấn đề phần mềm, và có thể trở nên dễ dàng hơn rất nhiều so với việc hoàn thiện công nghệ hình ảnh cần cho việc khởi tạo mẫu gốc. Các lập trình viên có thể dễ dàng thử nghiệm các mẹo như tăng số neuron trong các vùng vỏ não khác nhau để xem nó ảnh hưởng thế nào tới hiệu năng.⁷ Họ cũng có thể nghiên cứu việc tối ưu hóa mã và tìm kiếm những mô hình tính toán đơn giản hơn, có thể bảo tồn được chức năng cơ bản của từng neuron riêng biệt hay các mạng lưới thần kinh nhỏ. Nếu điều kiện tiên quyết cuối cùng về công nghệ là quét hoặc dịch, với công suất tính toán đã tương đối dư thừa, có thể ta sẽ không phải chú ý quá nhiều tới hiệu năng triển khai trong giai đoạn phát triển, và có thể có các cơ hội dễ dàng để tiết kiệm hiệu năng tính toán. (Việc tái tổ chức mang tính kiến trúc căn bản hơn cũng là khả thi, nhưng nó

đưa chúng đi lệch khỏi con đường giả lập và lẩn sân sang lãnh thổ AI).

Một cách khác để cải thiện cơ sở mã khi đã tạo ra được bản giả lập đầu tiên là quét các bộ não bổ sung có những kỹ năng/tài năng khác biệt hoặc nổi bật. Tăng trưởng năng suất cũng có thể xảy ra như một hệ quả của việc ứng dụng các cấu trúc tổ chức và sơ đồ luồng công việc vào các thuộc tính có một không hai của trí tuệ kỹ thuật số. Vì trong nền kinh tế của loài người chưa từng có tiền lệ về một công nhân có thể được sao chép, thiết lập lại, vận hành ở những tốc độ khác nhau, v.v. nên những người quản lý thể hệ công nhân giả lập đầu tiên sẽ tìm thấy rất nhiều không gian sáng tạo để đưa ra các phương pháp thực hành quản lý.

Sau giai đoạn lao dốc ban đầu, khi việc giả lập hoàn chỉnh não bộ con người trở nên khả thi, sức chống đối có thể tăng trở lại. Sớm muộn, những điểm hiệu năng triển khai kém hiển nhiên nhất sẽ được tối ưu hóa hết, những biến thể thuật toán hứa hẹn nhất sẽ được thử nghiệm và những cơ hội dễ dàng nhất cho việc sáng tạo tổ chức sẽ được khai thác. Thư viện mẫu sẽ được mở rộng sao cho việc lấy thêm các bản quét não sẽ mang lại lợi ích không đáng kể so với làm việc trên các mẫu hiện có. Vì một mẫu có thể được nhân bản, mỗi bản sao có thể được huấn luyện riêng trong những lĩnh vực khác nhau, và việc này sẽ được thực hiện với tốc độ điện tử, nên số lượng bộ não cần quét để có được hầu hết những tăng ích kinh tế tiềm năng sẽ rất nhỏ, thậm chí chỉ cần đến một bộ não mà thôi.

Một nguyên nhân tiềm tàng khác để sức chống đối tăng lên là khả năng của việc các bản giả lập hay những người ủng hộ sinh học của chúng sẽ tổ chức ủng hộ các quy định hạn chế việc sử dụng công

nhân giả lập, sao chép giả lập, cấm một số loại thí nghiệm với trí óc kỹ thuật số, thể chế hóa quyền công nhân và lương tối thiểu cho các công nhân giả lập, cùng nhiều hoạt động khác. Tuy nhiên, việc các diễn biến chính trị có thể đi theo hướng ngược lại, đóng góp vào sự suy giảm sức chống đối cũng có xác suất tương đương. Điều này có thể xảy ra nếu rào cản ban đầu trong việc sử dụng nhân công giả lập nhường đường cho hoạt động khai thác tự do khi cạnh tranh trở nên khốc liệt hơn và chi phí kinh tế và chiến lược của việc chiếm cứ những đỉnh cao trở nên rõ ràng.

Đối với trí tuệ nhân tạo (trí tuệ máy phi mô phỏng), khó khăn của việc nâng cấp một hệ thống từ cấp độ con người lên cấp độ siêu nhân bằng các biện pháp cải tiến thuật toán phụ thuộc vào các thuộc tính của hệ thống cụ thể. Những kiến trúc khác nhau có thể có sức chống đối rất khác nhau.

Trong một số trường hợp, sức chống đối có thể vô cùng thấp. Chẳng hạn, nếu AI cấp độ con người bị chậm trễ vì một sự thiếu hiểu cốt lõi nào đó vẫn lẫn lộn các lập trình viên, rồi sau đó khi bước đột phá cuối cùng xảy đến, AI có thể nhảy vọt từ mức thấp lên một mức cao hơn rất nhiều so với loài người mà không cần qua các bước trung gian. Một tình huống khác mà trong đó, sức chống đối có thể trở nên siêu thấp là khi một hệ thống AI có thể có được năng lực trí tuệ thông qua hai chế độ xử lý khác nhau. Để minh họa khả năng này, giả sử một AI được cấu thành từ hai tiểu hệ thống, một có các kỹ thuật xử lý tình huống có tính đặc thù lĩnh vực, và một có khả năng lý luận phổ thông. Khả năng lúc này, khi hệ thống thứ hai tiếp tục duy trì dưới một ngưỡng năng lực nhất định, nó sẽ không đóng góp gì cho hiệu năng chung của toàn hệ thống vì các giải pháp mà nó

đưa ra luôn thua kém những giải pháp của tiểu hệ thống đặc thù. Bây giờ, giả sử rằng một định lượng nhỏ năng lực tối ưu được đưa vào tiểu hệ thống phổ thông và khiến cho năng lực của tiểu hệ thống này gia tăng nhanh chóng. Đầu tiên, chúng ta không quan sát được sự tăng trưởng hiệu năng tổng thể của hệ thống, điều đó cho thấy sức chống đối còn cao. Sau đó, khi năng lực của tiểu hệ thống phổ thông vượt qua ngưỡng mà ở đó các giải pháp của nó bắt đầu chiến thắng các giải pháp của tiểu hệ thống đặc thù, hiệu năng tổng thể của hệ thống bất ngờ cải thiện với tốc độ nhanh chóng giống như tiểu hệ thống phổ thông, thậm chí khi định lượng công suất tối ưu hóa được đưa vào là không đổi. Lúc này, sức chống đối của hệ thống đã lao dốc.

Cũng có thể xu hướng nhìn nhận trí tuệ từ quan điểm dĩ nhân vi trung sẽ khiến chúng ta đánh giá thấp những cải thiện trong các hệ thống thấp hơn con người, và vì thế đánh giá cao sức chống đối. Eliezer Yudkowsky, nhà lý thuyết AI đã viết nhiều về tương lai của trí tuệ máy, phát biểu quan điểm này như sau (Xem thêm hình 8):

AI có thể thực hiện một bước nhảy đột ngột rõ ràng về trí tuệ là hoàn toàn do thuyết nhân hình, hay xu hướng của con người nghĩ về “chàng ngốc” và “Einstein” như những thái cực trên một thang đo trí tuệ, thay vì những điểm cận kề bất khả phân biệt trên một thang trí não nói chung. Bất cứ thứ gì ngốc hơn một chàng ngốc có thể hiển hiện với chúng ta đơn giản như là một kẻ “ngốc” mà thôi. Ta có thể hình dung “mũi tên AI” nhanh chóng vươn lên trên thang đo trí tuệ, vượt qua chuột và tinh tinh, khi mà AI vẫn là những tên “ngốc” vì chúng không nói năng trôi chảy hay viết được các bài báo khoa học. Sau đó, mũi tên AI này đi qua khoảng cách nhỏ nhoi giữa “ngoại-

ngốc” (infra- idiot) lên “ngoại-Einstein” (ultra-Einstein) chỉ trong một tháng hoặc tương đương.⁸

Những nhận xét này kết luận rằng khó mà dự báo được việc thực hiện các cải thiện thuật toán sẽ khó khăn đến mức nào trên cá thể AI đầu tiên đạt được trí tuệ tổng quát tương đương cấp độ con người. Ít nhất cũng có một số hoàn cảnh khả thể mà trong đó sức chống đối thuật toán là thấp. Nhưng thậm chí nếu sức chống đối thuật toán là rất cao thì nó cũng không làm cho sức chống đối tổng thể của AI đang được nghiên cứu bị giảm đi, vì ta có thể dễ dàng gia tăng trí tuệ của hệ thống theo những cách khác ngoài cải thiện thuật toán. Có hai yếu tố khác có thể được cải thiện: nội dung và phần cứng.



Hình 8: Một thang đo ít dựa vào thuyết nhân hình hơn? Khoảng cách giữa một kẻ ngốc và một người thông minh xem ra lớn hơn theo quan điểm dĩ nhân vi trung, nhưng từ một quan điểm ít thiên cận hơn, hai người này có trí óc gần như không thể phân biệt được.⁹ Chế tạo được một trí tuệ máy có mức độ thông minh tương đương với một chàng ngốc gần như chắc chắn là lâu hơn và khó khăn hơn so với việc cải thiện một hệ thống như vậy để trở nên thông minh hơn bất kỳ ai.

Trước hết hãy xem xét các biện pháp cải thiện nội dung. “Nội dung” ở đây được hiểu là những thành phần trong tổng thể phần mềm của một hệ thống không góp phần tạo ra kiến trúc thuật toán lõi của hệ thống đó. Chẳng hạn, nội dung có thể bao gồm các cơ sở dữ liệu đối tượng tri giác được lưu trữ, các thư viện kỹ năng chuyên biệt

và các kho kiến thức khai báo (declarative knowledge). Đối với nhiều loại hệ thống, sự phân biệt giữa kiến trúc thuật toán và nội dung rất không rõ nét; mặc dù vậy, nó vẫn sẽ là một cách tạm dùng được để chỉ ra một nguồn quan trọng tiềm tàng giúp tăng ích về năng lực trong một trí tuệ máy. Ý tưởng này về cơ bản có thể được diễn đạt rằng năng lực giải quyết vấn đề của một hệ thống có thể được tăng cường không chỉ bằng cách làm cho hệ thống thông minh hơn, mà còn bằng cách mở rộng những gì mà hệ thống đã biết.

Hãy xem xét một hệ thống AI hiện đại như TextRunner (một dự án nghiên cứu tại Đại học Washington) hoặc Watson của IBM (hệ thống đã chiến thắng trò chơi truyền hình Jeopardy!). Những hệ thống này có thể kết xuất những mảnh thông tin ngữ nghĩa nhất định bằng cách phân tích văn bản. Tuy không hiểu những gì đọc được theo cùng một ý nghĩa hay ở cùng mức độ như con người, nhưng chúng vẫn có thể kết xuất được một lượng thông tin đáng kể từ ngôn ngữ tự nhiên, và sử dụng thông tin này để đưa ra những suy luận đơn giản và trả lời câu hỏi. Chúng có thể học từ kinh nghiệm, tạo ra những phép biểu diễn khái quát hơn của một khái niệm khi gặp những trường hợp sử dụng bổ sung của khái niệm đó. Chúng được thiết kế để vận hành trong chế độ không cần giám sát trong phần lớn thời gian (chẳng hạn như học được cấu trúc ẩn trong dữ liệu không gán nhãn khi không có tín hiệu về lỗi hay phần thưởng mà không cần đến sự hướng dẫn của con người), để có tốc độ cao và khả năng mở rộng. Ví dụ, TextRunner làm việc với một tập hợp 500 triệu trang web.¹⁰

Bây giờ, hãy hình dung một hậu duệ xa của một hệ thống như vậy có được khả năng đọc và hiểu chỉ như một đứa trẻ 10 tuổi,

nhưng tốc độ đọc như TextRunner. (Đây có thể là một vấn đề AI toàn diện). Như vậy, chúng ta đang hình dung ra một hệ thống suy nghĩ nhanh hơn và có bộ nhớ tốt hơn một người trưởng thành, nhưng biết ít hơn nhiều, và hiệu ứng tổng thể của việc này có thể là hệ thống gần tương đương con người về năng lực giải quyết vấn đề phổ thông. Tuy nhiên, sức chống đối nội dung của nó là rất thấp – đủ thấp để khiến quá trình cắt cánh diễn ra sớm hơn. Trong vòng vài tuần, hệ thống có thể đọc và làm chủ được toàn bộ nội dung có trong Thư viện Quốc hội. Giờ đây, hệ thống này đã biết nhiều hơn bất kỳ người nào và có tốc độ suy nghĩ tăng lên đáng kể; nó đã (ít nhất) trở thành một siêu trí tuệ yếu.

Vì vậy, một hệ thống có thể gia tăng đáng kể năng lực trí tuệ hiệu dụng bằng cách hấp thụ nội dung đã được tạo ra từ trước, tích lũy qua nhiều thế kỷ khoa học và văn minh của loài người, chẳng hạn, bằng cách đọc trên Internet. Nếu một AI đạt đến trình độ con người mà trước đó không tiếp cận với tư liệu này hay không thể tiêu hóa được nó, thì khi đó sức chống đối chung của AI này sẽ thấp ngay cả nếu việc cải thiện kiến trúc thuật toán của nó là khó khăn.

Sức chống đối nội dung cũng là một khái niệm phù hợp với các bản giả lập. Một bản giả lập tốc độ cao có ưu thế không chỉ vì nó có thể hoàn thành cùng những nhiệm vụ như con người sinh học với tốc độ nhanh hơn, mà còn vì nó có thể tích lũy nhiều hơn những nội dung đúng lúc, ví dụ như những kỹ năng và kinh nghiệm liên quan đến một nhiệm vụ nhất định. Tuy nhiên, để khai thác được toàn bộ tiềm năng của việc tích lũy nhanh chóng nội dung, một hệ thống cần có dung lượng bộ nhớ lớn tương ứng. Đọc hết cả thư viện sẽ không có nhiều ý nghĩa nếu đọc trước quên sau. Trong khi một hệ thống AI

có thể có đủ dung lượng bộ nhớ, các bản giả lập lại thừa hưởng một vài hạn chế về dung lượng từ những người làm mẫu. Do đó, chúng có thể cần tăng cường kiến trúc để có khả năng học tập không giới hạn.

Từ đầu đến giờ, chúng ta đã xem xét sức chống đối về kiến trúc và nội dung – nghĩa là mức độ khó khăn của việc cải thiện phần mềm của một trí tuệ máy đã đạt cấp độ con người. Bây giờ hãy cùng xem xét phương pháp thứ ba để nâng cao hiệu năng của trí tuệ máy: cải thiện phần cứng. Vậy sức chống đối với những cải thiện về phần cứng là gì?

Bắt đầu với phần mềm thông minh (giả lập hoặc AI), ta có thể khuếch đại trí tuệ tập thể chỉ đơn giản bằng cách sử dụng thêm máy tính để chạy nhiều bản sao của phần mềm hơn.¹¹ Ta cũng có thể khuếch đại trí tuệ tốc độ bằng cách chuyển chương trình phần mềm sang một máy tính nhanh hơn. Tùy thuộc vào mức độ mà chương trình có thể chạy song song, trí tuệ tốc độ cũng có thể được khuếch đại bằng cách chạy chương trình trên nhiều bộ xử lý hơn. Việc này rất khả thi với các bản mô phỏng có kiến trúc được song song hóa ở mức độ cao; nhưng nhiều chương trình AI cũng có các chương trình con quan trọng có thể thu được lợi ích từ việc song song hóa quy mô lớn. Việc khuếch đại trí tuệ chất lượng cũng có thể được thực hiện bằng cách gia tăng công suất tính toán, tuy nhiên đây lại không phải là việc đơn giản.¹²

Sức chống đối với việc khuếch đại trí tuệ tập thể hoặc trí tuệ tốc độ (và có thể là trí tuệ chất lượng) trong một hệ thống với phần mềm cấp độ con người do đó có thể là thấp. Khó khăn duy nhất ở đây là truy cập được công suất tính toán bổ sung. Có một số cách để một

hệ thống mở rộng cơ sở phần cứng của mình, và mỗi cách phù hợp với những thang thời gian khác nhau.

Trong ngắn hạn, công suất tính toán phải mở rộng gần như tuyến tính với ngân quỹ: ngân quỹ tăng gấp đôi mua được số lượng máy tính gấp đôi, cho phép số bản sao phần mềm nhiều gấp đôi chạy cùng một lúc. Sự nổi lên của các dịch vụ điện toán đám mây mang lại cho một dự án thêm lựa chọn mở rộng tài nguyên tính toán mà không phải chờ được bàn giao và lắp đặt máy tính mới, mặc dù các quan ngại về tính bảo mật có thể khiến mọi người thiên về việc sử dụng máy tính tại chỗ. (Trong một số kịch bản nhất định, công suất tính toán có thể có được bằng các biện pháp khác, như bằng cách kiểm soát các botnet.¹³) Hệ thống ban đầu sử dụng bao nhiêu công suất tính toán sẽ quyết định độ khó của việc mở rộng hệ thống đó với một hệ số cho trước. Một hệ thống ban đầu chạy trên một máy PC có thể được mở rộng với hệ số lên đến hàng ngàn chỉ với một triệu đô la, còn một chương trình chạy trên một siêu máy tính sẽ tốn kém hơn nhiều để mở rộng.

Trong giai đoạn dài hạn hơn một chút, chi phí mua thêm phần cứng có thể tăng lên khi ngày càng nhiều dung lượng đã được lắp đặt trên thế giới được sử dụng để chạy các trí não kỹ thuật số. Ví dụ, trong một kịch bản giả lập cạnh tranh theo thị trường, chi phí vận hành thêm một bản sao của một bản giả lập có thể tăng lên xấp xỉ mức thu nhập do bản sao đó tạo ra, vì các nhà đầu tư nâng giá hạ tầng tính toán hiện có cho phù hợp với mức hoàn vốn mà họ kỳ vọng nhận được từ khoản đầu tư của mình (mặc dù nếu chỉ có một dự án làm chủ được công nghệ, dự án này sẽ có thể có được một mức độ

độc quyền mua nào đó trong thị trường công suất tính toán và vì thế sẽ trả mức giá thấp hơn.)

Trong một thang thời gian dài hơn nữa, nguồn cung công suất tính toán sẽ phát triển khi dung lượng mới được lắp đặt. Lượng cầu tăng vọt sẽ kích thích mạnh mẽ việc sản xuất trong các nhà máy bán dẫn hiện có và thúc đẩy việc xây dựng những nhà máy mới. (Sự tăng hiệu năng một lần, có thể lên đến hai hoặc ba cấp khuếch đại, có thể đạt được bằng cách sử dụng những bộ vi xử lý đã tùy biến.¹⁴) Trên hết, làn sóng cải tiến công nghệ đang dâng trào sẽ đổ khối lượng công suất tính toán ngày càng tăng vào tua-bin của những chiếc máy tư duy. Về mặt lịch sử, tốc độ cải tiến công nghệ tính toán đã được mô tả trong định luật Moore nổi tiếng, mà một trong những biến thể của định luật này phát biểu rằng công suất tính toán trên một đô la tăng gấp đôi sau mỗi 18 tháng hoặc xấp xỉ.¹⁵ Mặc dù ta không thể cho rằng tốc độ cải thiện này áp dụng được cho việc phát triển trí tuệ máy cấp độ con người, nhưng cho tới khi các giới hạn vật lý chưa bị chinh phục thì vẫn sẽ còn có không gian cho những tiến bộ trong công nghệ tính toán.

Vì thế, ta có lý do để kỳ vọng rằng sức chống đối phần cứng sẽ không quá cao. Mua thêm công suất tính toán cho hệ thống khi nó chứng minh được khả năng có được trí tuệ cấp độ con người có thể dễ dàng bổ sung vài cấp khuếch đại nữa cho công suất tính toán (tùy thuộc vào mức độ nghèo nàn về phần cứng của dự án trước khi mở rộng). Tùy biến chip có thể bổ sung thêm một hay hai cấp khuếch đại nữa. Những phương pháp mở rộng cơ sở phần cứng khác như xây dựng thêm nhà máy và đẩy xa hơn biên giới của công nghệ máy tính sẽ tốn thời gian hơn – thường là vài năm, mặc dù độ trễ này có thể

giảm mạnh một khi siêu trí tuệ máy cách mạng hóa quá trình sản xuất và phát triển công nghệ.

Tóm lại, chúng ta có thể nói về xác suất của một mái hiên phần cứng (hardware overhang): khi phần mềm cấp độ con người được tạo ra, công suất tính toán có thể đã đầy đủ từ trước để chạy rất nhiều bản sao của phần mềm đó với tốc độ cao. Sức chống đối phần mềm, như đã trao đổi ở trên, là khó đánh giá hơn nhưng có thể còn thấp hơn sức chống đối phần cứng. Cụ thể, có thể có mái hiên phần mềm (software overhang) dưới dạng nội dung đã có sẵn (như Internet) trở nên sẵn có cho hệ thống khi nó đạt đến trình độ ngang với con người. Mái hiên thuật toán (algorithm overhang) – những phần tăng cường bằng thuật toán được thiết kế sẵn – cũng là khả thể nhưng có lẽ với xác suất thấp hơn. Các cải tiến phần mềm (kể cả thuật toán hay nội dung) có thể tăng thêm vài cấp khuếch đại cho những tăng ích hiệu năng tiềm tàng mà có thể tiếp cận được tương đối dễ dàng một khi trí óc số ngang với con người, ngoài những tăng ích hiệu năng có được bằng cách sử dụng phần cứng nhiều hoặc tốt hơn.

Công suất tối ưu hóa và sự bùng nổ

Sau khi nghiên cứu vấn đề sức chống đối, chúng ta cần chuyển sang nửa còn lại của phương trình giản lược: công suất tối ưu hóa. Hãy nhớ lại: Tốc độ thay đổi trí tuệ = Công suất tối ưu hóa/Sức chống đối. Như được thể hiện trong dạng giản ước này, quá trình cất cánh nhanh không đòi hỏi sức chống đối trong giai đoạn chuyển đổi phải thấp. Một quá trình cất cánh nhanh có thể xảy ra nếu sức chống đối không thay đổi hoặc thậm chí tăng một cách khiêm tốn, với điều kiện công suất tối ưu hóa áp dụng để cải thiện hiệu năng của hệ

thống tăng với tốc độ đủ nhanh. Như chúng ta sẽ thấy sau đây, có một số cơ sở chắc chắn để nghĩ rằng công suất tối ưu hóa được áp dụng sẽ tăng lên trong quá trình chuyển đổi, ít nhất là khi không có những biện pháp cố ý để ngăn không cho điều đó xảy ra.

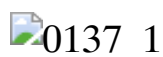
Chúng ta có thể phân biệt hai pha. Pha thứ nhất bắt đầu với giai đoạn khởi đầu của quá trình cất cánh, khi hệ thống đạt đến đường cơ sở con người đối với trí tuệ cá nhân. Khi năng lực của hệ thống tiếp tục tăng lên, nó có thể sử dụng một phần hoặc toàn bộ năng lực đó để tự cải thiện (hoặc để thiết kế một hệ thống kế tục – mà với những mục tiêu hiện tại, cũng để đạt được sự cải thiện). Tuy nhiên, phần lớn công suất tối ưu hóa được áp dụng cho hệ thống vẫn đến từ bên ngoài, có thể là từ quá trình làm việc của các lập trình viên hay kỹ sư làm việc trong dự án hay từ những công việc tương tự được thực hiện ở thế giới bên ngoài có thể được dự án tiếp nhận và sử dụng.¹⁶ Nếu pha này kéo dài tương đối lâu, chúng ta có thể dự đoán rằng lượng công suất tối ưu hóa được áp dụng cho hệ thống sẽ tăng lên. Đầu vào từ bên trong dự án và từ thế giới bên ngoài có thể sẽ tăng lên khi những hứa hẹn của cách tiếp cận này bắt đầu trở thành hiện thực. Các nhà nghiên cứu có

thể làm việc chăm chỉ hơn, các nhà nghiên cứu khác có thể được chiêu mộ, và công suất tính toán cũng có thể được mua thêm để đẩy nhanh tiến độ. Sự gia tăng này có thể trở nên đặc biệt kịch tính nếu việc phát triển trí tuệ máy cấp độ con người làm thế giới bất ngờ, mà khi ấy, dự án nghiên cứu mà trước đó từng có quy mô nhỏ có thể bất ngờ trở thành tiêu điểm của những nỗ lực nghiên cứu phát triển mạnh mẽ trên khắp thế giới (mặc dù một vài nỗ lực trong đó dễ được đưa vào các dự án cạnh tranh lẫn nhau).

Pha tăng trưởng thứ hai sẽ bắt đầu nếu vào một thời điểm nào đó hệ thống có được năng lực cao tới mức phần lớn công suất tối ưu hóa đưa vào hệ thống xuất phát từ chính nó (được đánh dấu bởi mức biến đổi gọi là “điểm quá độ” trong Hình 7). Điều đó sẽ làm thay đổi động lực một cách cơ bản, vì bất cứ sự gia tăng nào của năng lực hệ thống giờ đây sẽ được chuyển dịch thành một sự gia tăng tương ứng của lượng công suất tối ưu hóa được áp dụng vào quá trình cải thiện hơn nữa của chính hệ thống. Nếu sức chống đối là không đổi, động lực hồi tiếp này sẽ tạo ra sự tăng trưởng theo cấp số nhân (Xem Khung 4). Hằng số nhân đôi phụ thuộc vào kịch bản này nhưng có thể rất ngắn – chỉ một vài giây trong một số kịch bản – nếu sự tăng trưởng diễn ra với tốc độ điện tử, điều có thể xảy ra như là hệ quả của những cải thiện thuật toán hoặc việc khai thác một mái hiên (overhang) về nội dung hoặc phần cứng.¹⁷ Sự tăng trưởng do các công trình vật lý tạo ra, ví dụ sản xuất các máy tính hoặc thiết bị mới, sẽ đòi hỏi thang thời gian dài hơn đôi chút (nhưng vẫn là một thang thời gian có thể là rất ngắn so với tốc độ tăng trưởng hiện nay của nền kinh tế thế giới).

Khung 4: Bàn về động lực của một sự bùng nổ trí tuệ

Chúng ta có thể viết tốc độ thay đổi trí tuệ thành tỷ lệ giữa công suất tối ưu hóa được áp dụng vào hệ thống và sức chống đối của hệ thống.



Lượng công suất tối ưu hóa hoạt động trên một hệ thống là tổng của bất kỳ lượng công suất tối ưu hóa nào do chính hệ thống đóng

góp và lượng công suất tối ưu hóa từ bên ngoài. Ví dụ, một AI hạt giống có thể được cải thiện nhờ một tổ hợp nỗ lực của chính nó và một nhóm lập trình viên, và có thể cả nỗ lực của cộng đồng các nhà nghiên cứu toàn cầu rộng lớn hơn làm nên những tiến bộ liên tục trong ngành bán dẫn, khoa học máy tính và các lĩnh vực liên quan khác.¹⁸

 0137_2

Một AI hạt giống khởi đầu với những năng lực nhận thức rất hạn chế. Vì vậy, vào giai đoạn bắt đầu,  0137_3 sẽ có giá trị nhỏ.¹⁹

Vậy còn  0137_4 và  thế giới thì sao?

Có những trường hợp mà trong đó một dự án đơn lẻ có nhiều năng lực thích hợp hơn so với toàn bộ phần còn lại của thế giới kết hợp lại – chẳng hạn, dự án Manhattan đã đưa một phần rất lớn các nhà vật lý giỏi nhất thế giới đến Los Alamos để chế tạo bom nguyên tử. Phổ biến hơn, bất kỳ dự án nào cũng chỉ sở hữu một phần nhỏ của tổng năng lực nghiên cứu phù hợp trên toàn thế giới. Nhưng ngay cả khi thế giới bên ngoài có tổng lượng năng lực nghiên cứu phù hợp lớn hơn bất cứ dự án nào,  0137_4 vẫn có thể cao hơn

 thế giới vì phần lớn năng lực của thế giới bên ngoài không tập trung vào hệ thống cụ thể đang được quan tâm. Nếu một dự án bắt đầu có vẻ hứa hẹn – điều sẽ xảy đến khi hệ thống vượt qua đường cơ sở con người nếu như không phải trước đó – nó có thể thu hút thêm đầu tư và làm tăng  0137_4. Nếu những thành quả của dự

án là của công chúng, $\mathcal{D}_{\text{thế giới}}$ cũng có thể tăng lên khi tiến bộ tạo cảm hứng cho những mối quan tâm lớn hơn với trí tuệ máy nói chung và khi các đại gia khác tìm cách gia nhập sân chơi. Do đó, trong pha chuyển đổi, công suất tối ưu hóa tổng được áp dụng để cải thiện một hệ thống nhận thức có lẽ sẽ tăng lên khi năng lực của hệ thống tăng lên.²⁰

Khi năng lực của hệ thống tăng lên, có thể sẽ có một thời điểm mà công suất tối ưu hóa do chính hệ thống tạo ra bắt đầu vượt qua công suất tối ưu hóa được đưa vào từ bên ngoài (trên toàn bộ các chiều kích đáng kể của quá trình cải thiện):

$$\mathcal{D}_{\text{hệ thống}} > \mathcal{D}_{\text{dự án}} + \mathcal{D}_{\text{thế giới}}$$

Điểm quá độ này là quan trọng, bởi lẽ khi vượt qua điểm này, việc tiếp tục cải thiện năng lực của hệ thống sẽ đóng góp mạnh mẽ vào sự gia tăng của tổng công suất tối ưu hóa được áp dụng để cải thiện hệ thống. Bằng cách đó, chúng ta bước vào một chế độ tự cải thiện đệ quy mạnh mẽ. Việc này dẫn đến sự phát triển bùng nổ của năng lực hệ thống dưới tương đối nhiều dạng khác nhau của đường biểu diễn sức chống đối.

Để minh họa, trước tiên hãy xem xét một kịch bản mà trong đó sức chống đối là hằng số, sao cho tốc độ tăng trí tuệ của một AI bằng công suất tối ưu hóa được áp dụng. Giả sử toàn bộ công suất tối ưu hóa được áp dụng đến từ chính AI, và AI dùng toàn bộ trí tuệ của

mình vào việc khuếch đại trí tuệ của chính nó, sao cho  0138_2²¹
Khi đó ta có:

$$\frac{dI}{dt} = \frac{I}{k}.$$

Giải phương trình vi phân đơn giản này cho ta hàm mũ:

$$I = Ae^{t/k}.$$

Nhưng sức chống đổi là hằng số chỉ là một trường hợp đặc biệt. Sức chống đổi có thể giảm xung quanh đường cơ sở con người, vì một hay nhiều yếu tố được nhắc đến trong những phần trước, và tiếp tục duy trì ở mức thấp quanh điểm quá độ và một khoảng nhất định phía trên (có lẽ cho tới khi hệ thống tiếp cận đến những giới hạn vật lý cơ bản). Giả sử, công suất tối ưu hóa áp dụng vào hệ thống gần

như không đổi (tức là $O_{\text{dự án}} + O_{\text{thế giới}} \approx C$) trước khi hệ thống có thể đóng góp đáng kể vào thiết kế riêng của nó, và điều này dẫn đến việc hệ thống nhân đôi năng lực sau mỗi 18 tháng. (Điều này có thể gần tương đồng với tốc độ cải thiện trong lịch sử từ định luật Moore kết hợp với những tiến bộ về phần mềm.²²) Tốc độ cải thiện này, nếu đạt được bằng công suất tối ưu hóa gần như không đổi, sẽ

kéo theo việc sức chống đối giảm xuống theo tỷ lệ nghịch với năng lực hệ thống:

$$\frac{dI}{dt} = \frac{c}{1/I} = cI.$$

Nếu sức chống đối tiếp tục giảm theo đồ thị hyperbol này thì khi AI đạt đến điểm quá độ, tổng công suất tối ưu hóa được áp dụng để cải thiện AI đã tăng gấp đôi. Khi đó ta có:

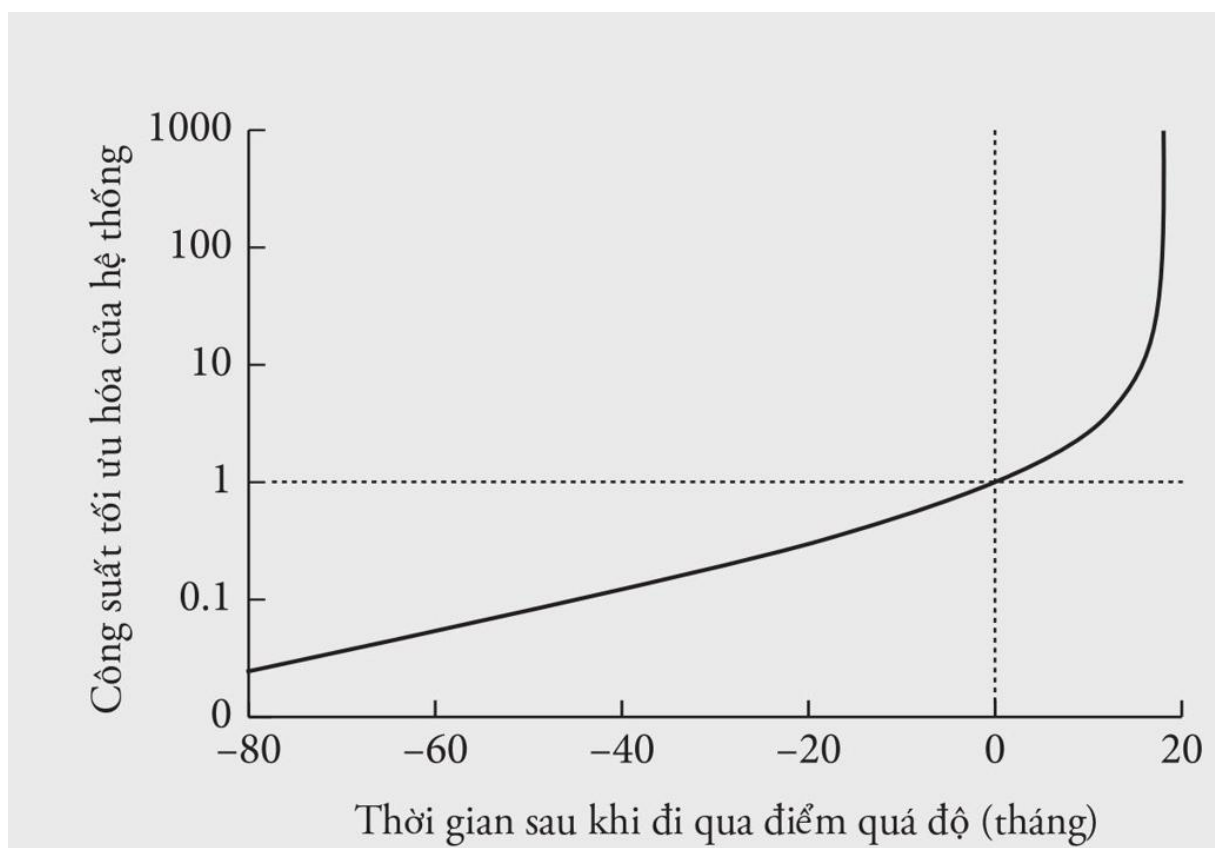
$$\frac{dI}{dt} = \frac{(c+I)}{1/I} = (c+I)I.$$

Lần nhân đôi thứ hai xảy ra 7,5 tháng sau đó. Trong 17,9 tháng, năng lực của hệ thống đã tăng lên một ngàn lần, và như vậy đã đạt đến siêu trí tuệ tốc độ (Hình 9).

Quỹ đạo tăng trưởng đặc biệt này có một điểm kỳ dị dương tại $t = 18$ tháng. Trên thực tế, giả định rằng sức chống đối là hằng số sẽ không còn giá trị khi hệ thống bắt đầu tiệm cận các giới hạn vật lý của quá trình xử lý thông tin, nếu như không phải là sớm hơn.

Hai kịch bản này chỉ có mục đích minh họa; nhiều quỹ đạo khác cũng có thể có, tùy vào hình dạng của đường biểu diễn sức chống đối. Kết luận ở đây đơn giản là vòng lặp hồi tiếp mạnh mẽ xuất hiện

quanh điểm quá độ rất có khả năng sẽ khiến quá trình cất cánh nhanh diễn ra nhanh hơn so với khi không có nó.



Hình 9: Đồ thị đơn giản của sự bùng nổ trí tuệ.

Có thể công suất tối ưu hóa được áp dụng sẽ tăng lên trong khi chuyển đổi: ban đầu bởi vì con người cố gắng hơn nữa để cải thiện một trí tuệ máy đang cho thấy một sự hứa hẹn đầy hấp dẫn, sau đó là do chính trí tuệ máy trở nên có khả năng thúc đẩy hơn nữa quá trình tiến bộ với một tốc độ “kỹ thuật số”. Việc này sẽ khiến quá trình cất cánh nhanh hoặc trung bình thật sự có khả năng xảy ra, ngay cả nếu sức chống đối là không đổi hoặc chỉ tăng lên đôi chút xung quanh đường cơ sở con người.²³

Mặc dù vậy, trong những phần trước, chúng ta đã thấy rằng có những yếu tố có thể dẫn đến sự sụt giảm lớn về sức chống đối quanh đường cơ sở năng lực con người. Chẳng hạn, những yếu tố này bao gồm khả năng nhanh chóng mở rộng phần cứng khi có được một trí óc phần mềm hiệu quả; khả năng cải thiện thuật toán; khả năng quét thêm các bộ não bổ sung (trong trường hợp giả lập hoàn chỉnh não bộ); và khả năng nhanh chóng tích hợp một lượng thông tin khổng lồ bằng cách tiêu hóa thông tin trên Internet (trong trường hợp trí tuệ nhân tạo).²⁴

Mặc dù đã có những quan sát này, nhưng hình dạng của đường biểu diễn sức chống đối trong khu vực liên quan vẫn chưa được đặc tả đầy đủ. Cụ thể, mức độ khó khăn của việc cải thiện chất lượng phần mềm trên một bản giả lập cấp độ con người hay một AI là chưa rõ ràng, và sự khó khăn của việc tăng công suất phần cứng sẵn có cho hệ thống cũng vậy. Vì hiện nay việc tăng công suất tính toán sẵn có cho một dự án nhỏ là tương đối dễ dàng, bằng cách tăng khoản chi cho công suất tính toán lên hàng ngàn lần hoặc chờ vài năm cho tới khi giá máy tính hạ xuống, trí tuệ máy đầu tiên đạt đến vạch mốc con người có thể sẽ được tạo ra từ một dự án lớn với những siêu máy tính đắt giá không thể nâng cấp với chi phí thấp, và có thể định luật Moore khi đó cũng không còn đúng nữa. Vì những lý do này, mặc dù một quá trình cất cánh trung bình hay nhanh có vẻ khả thể hơn, ta cũng cũng không thể loại trừ khả năng xảy ra một quá trình cất cánh chậm.²⁵

Chương 5

Lợi thế cạnh tranh quyết định

Một câu hỏi khác nhưng có liên quan đến câu hỏi về động lực là liệu sẽ có một hay nhiều thế lực siêu trí tuệ? Một sự bùng nổ trí tuệ có thể đẩy một dự án đi xa hơn các dự án khác, đến mức khiến nó có thể độc chiếm tương lai hay không? Hay liệu sự tiến bộ có đồng dạng hơn, trải đều trên diện rộng, với nhiều dự án tham gia nhưng không dự án nào có được vị thế dẫn đầu lâu dài và áp đảo?

Chương trước đã phân tích một tham số chính yếu khi xác định kích thước của khoảng cách có thể mở ra giữa thế lực đi đầu và những đối thủ cạnh tranh gần nhất – cụ thể là tốc độ chuyển đổi từ con người sang trí tuệ siêu nhân mạnh mẽ. Điều này khuyến nghị một phân tích mở màn. Nếu quá trình cất cánh là nhanh (hoàn thành trong vài giờ, vài ngày hay vài tuần), ít có khả năng hai dự án độc lập có thể đồng thời cất cánh: gần như chắc chắn rằng dự án đầu tiên đã hoàn thành quá trình cất cánh của mình trước khi bất cứ dự án nào khác bắt đầu. Nếu quá trình cất cánh là chậm (trải dài nhiều năm hoặc thập kỷ), thì có thể nhiều dự án sẽ cùng cất cánh đồng thời, sao cho ngay cả khi các dự án có thể có được năng lực to lớn vào giai đoạn cuối quá trình chuyển đổi, thì vẫn không có thời điểm nào mà ở đó bất cứ dự án nào có thể vượt quá xa các dự án khác để chiếm vị trí dẫn đầu. Quá trình cất cánh tốc độ trung bình nằm ở giữa, với cả

hai điều kiện đều có thể: có thể có hoặc không có nhiều hơn một dự án đang cất cánh cùng một lúc.¹

Liệu một trí tuệ máy có vượt xa trong cuộc cạnh tranh tới mức có được một lợi thế cạnh tranh quyết định – tức là một mức độ phát triển về công nghệ và các lĩnh vực khác đủ để cho phép nó thống trị thế giới hoàn toàn hay không? Nếu một dự án có được lợi thế cạnh tranh quyết định, liệu nó có sử dụng ưu thế này để đàn áp các đối thủ cạnh tranh và tạo ra một thể đơn nhất (singleton – một trật tự thế giới chỉ có một tổ chức ra quyết định duy nhất ở cấp độ toàn cầu) hay không? Và nếu có một dự án chiến thắng, nó sẽ “lớn” đến mức nào (không phải là về kích thước vật lý hay ngân sách mà về việc mong muốn của bao nhiêu người sẽ kiểm soát thiết kế của nó)? Chúng ta sẽ lần lượt xem xét từng vấn đề này.

Người dẫn đầu có giành lợi thế cạnh tranh quyết định?

Một yếu tố ảnh hưởng đến khoảng cách giữa người dẫn đầu và những người theo sau là tốc độ khuếch tán của bất kỳ điều gì tạo cho người dẫn đầu một ưu thế cạnh tranh cho dù thứ đó là gì đi nữa. Một người dẫn đầu có thể thấy rằng khó mà tạo ra và duy trì được một khoảng cách lớn nếu những người đi sau có thể dễ dàng sao chép ý tưởng và sáng tạo của mình. Sự bắt chước tạo ra một luồng “gió ngược” có hại cho người dẫn đầu và có lợi cho những kẻ theo sau, đặc biệt nếu quyền sở hữu trí tuệ chỉ được bảo vệ một cách hời hợt. Một người dẫn đầu có thể đặc biệt dễ trở thành đối tượng trưng dụng tài sản, đánh thuế hoặc bị hủy hoại dưới áp lực của quy định chống độc quyền.

Tuy nhiên sẽ là sai lầm nếu giả sử rằng cơn gió ngược này phải tăng đơn điệu theo khoảng cách giữa người dẫn đầu và những

người đi sau. Cũng giống như một vận động viên đua xe đạp bị rút lại quá xa trong cuộc đua và không còn được những vận động viên chạy trước chắn gió, những người đi sau về công nghệ bị rút lại đủ xa cũng sẽ thấy khó nắm bắt được những tiến bộ đã được làm ra ở phía trước.² Khoảng cách về hiểu biết và năng lực có thể trở nên quá lớn. Người dẫn đầu có thể đã chuyển sang nền tảng công nghệ tiên tiến hơn, khiến những sáng tạo tiếp theo không thể chuyển giao được sang những nền tảng thô sơ mà những người đi sau đang sử dụng. Một người đi đầu nổi trội có thể có khả năng ngăn chặn sự rò rỉ thông tin từ các chương trình nghiên cứu và những cơ sở nhạy cảm của mình, hoặc phá hoại những nỗ lực của đối thủ cạnh tranh trong việc phát triển những năng lực tiên tiến.

Nếu người đi trước là một hệ thống AI, nó có thể có những thuộc tính giúp tự mở rộng năng lực của mình dễ dàng hơn trong khi giảm tốc độ khuếch tán. Trong những tổ chức do con người điều hành, các lợi thế kinh tế nhờ quy mô (economies of scale) bị mất tác dụng do vấn đề hiệu năng kém mang tính quan liêu và các vấn đề tổ chức, bao gồm cả những khó khăn trong việc bảo vệ bí mật thương mại.³ Những vấn đề này được cho là sẽ hạn chế sự tăng trưởng của một dự án trí tuệ máy, chừng nào nó còn do con người điều hành. Tuy nhiên, một hệ thống AI lại có thể tránh được một số điểm phi kinh tế do quy mô (scale diseconomies) này vì các module của AI (trái ngược với công nhân) không có những ưu tiên cá nhân khác với những ưu tiên của cả hệ thống. Do đó, hệ thống AI có thể tránh được một phần lớn các điểm kém hiệu quả nảy sinh từ các vấn đề tổ chức trong một doanh nghiệp của con người. Ưu thế này (việc có những thành viên trung thành đến hoàn hảo) cũng giúp cho một hệ thống AI

dễ theo đuổi hơn những mục tiêu bí mật dài hạn. Một hệ thống AI sẽ không có các nhân viên giận dữ sẵn sàng trở thành con mồi cho đối thủ cạnh tranh hay bị mua chuộc để trở thành kẻ bán thông tin.⁴

Chúng ta có thể cảm nhận được sự phân bố của những khoảng cách có thể có trong các thời kỳ phát triển bằng cách xem xét một số ví dụ lịch sử (xem Khung 5). Xem ra độ chênh lệch khoảng một vài tháng đến một vài năm là khá tiêu biểu cho những dự án chiến lược công nghệ đáng kể.

Có thể sự toàn cầu hóa và quá trình giám sát được gia tăng sẽ làm giảm khoảng tụt hậu tiêu biểu giữa các dự án công nghệ đang cạnh tranh nhau. Mặc dù vậy, những khoảng tụt hậu trung bình này chỉ có thể rút ngắn xuống đến một giới hạn dưới nào đó (khi thiếu vắng sự điều phối có chủ đích).⁵ Ngay cả khi không có động lực dẫn đến các hiệu ứng quả cầu tuyết, một số dự án cũng vẫn sẽ tình cờ có được nhân viên nghiên cứu, ban lãnh đạo và hạ tầng tốt hơn, hoặc tình cờ va phải những ý tưởng hay hơn. Nếu hai dự án theo đuổi hai cách tiếp cận khác nhau mà một trong số đó có hiệu quả hơn, dự án còn lại có thể sẽ phải mất nhiều tháng để chuyển sang cách tiếp cận vượt trội, ngay cả khi có thể giám sát chặt chẽ những gì mà dự án dẫn trước đang làm.

Khung 5 Những cuộc đua công nghệ: một vài ví dụ lịch sử

Trong suốt chiều dài lịch sử, đã có sự gia tăng về tốc độ khuếch tán kiến thức và công nghệ trên thế giới, và hệ quả là khoảng cách về thời gian giữa những người đi đầu về công nghệ và những người theo sau gần nhất đã bị rút ngắn.

Trung Quốc đã tìm cách duy trì vị thế độc quyền trong ngành sản xuất lụa trong gần 2.000 năm. Kết quả khảo cổ cho thấy việc sản xuất

có thể đã bắt đầu khoảng năm 3000 TCN, hay thậm chí sớm hơn.⁶ Ngành trồng dâu nuôi tằm là một bí mật được cất giữ cẩn thận. Hành vi tiết lộ kỹ thuật, cũng như việc xuất khẩu tằm hay trứng tằm ra ngoài lãnh thổ Trung Quốc bị khép vào tội tử hình. Những người La Mã, mặc dù lụa tơ tằm được nhập khẩu vào đế quốc này có giá rất cao, vẫn không học được nghệ thuật sản xuất lụa tơ tằm. Cho tới năm 300, một đoàn thám hiểm Nhật Bản đã tìm cách bắt được một ít trứng tằm cùng với ba thiếu nữ Trung Quốc. Những thiếu nữ này bị buộc phải tiết lộ nghệ thuật lụa tơ tằm cho những kẻ bắt cóc mình.⁷ Thành Byzantium cũng gia nhập câu lạc bộ các nhà sản xuất vào năm 522. Sự tụt hậu cũng xuất hiện trong câu chuyện về việc sản xuất đồ sứ. Người ta đã hành nghề này ở Trung Quốc từ đời Tống khoảng năm 600 (và có thể đồ sứ đã được sử dụng từ những năm 200), nhưng phải đến thế kỷ 18, người châu Âu mới có thể làm được thành thạo.⁸ Xe có bánh xuất hiện ở một số địa điểm ở châu Âu và vùng Lưỡng Hà vào khoảng năm 3500 TCN, nhưng cũng phải đến thời kỳ hậu Colombo mới đến được nước Mỹ.⁹ Trên quy mô lớn hơn, loài người cần 10 ngàn năm để lan tỏa hầu hết địa cầu, cuộc Cách mạng Nông nghiệp cần hàng ngàn năm, Cách mạng Công nghiệp chỉ cần vài trăm năm và một cuộc Cách mạng thông tin có thể nói đã lan tỏa trên toàn cầu chỉ trong khoảng thời gian vài thập kỷ, mặc dù tất nhiên là những sự chuyển giao này không hẳn là có mức độ sâu sắc tương đương nhau. (Trò chơi video Dance Dance Revolution lan truyền từ Nhật sang Châu Âu và Bắc Mỹ chỉ trong vòng một năm!)

Vấn đề cạnh tranh công nghệ đã được nghiên cứu khá kỹ lưỡng, đặc biệt trong bối cảnh các cuộc chạy đua bằng sáng chế và vũ

trang.¹⁰ Việc xem xét lại những tài liệu này nằm ngoài phạm vi nghiên cứu chúng tôi ở đây. Tuy nhiên, nghiên cứu một số ví dụ của những cuộc chạy đua công nghệ quan trọng mang tính chiến lược trong thế kỷ 20 cũng cho biết nhiều điều (xem Bảng 7).

Liên quan đến sáu công nghệ trong bảng này, các công nghệ được các cường quốc đối địch coi là có tầm quan trọng chiến lược vì ý nghĩa quân sự hay biểu tượng của chúng, sự tụt hậu giữa nước dẫn đầu và nước đi sau gần nhất (theo đánh giá rất sát) tương ứng là 49 tháng, 4 tháng, 1 tháng, 4 tháng và 60 tháng – lâu hơn khoảng thời gian của quá trình cất nhanh và ngắn hơn khoảng thời gian của quá trình cất cánh chậm.¹¹ Trong nhiều trường hợp, dự án của nước đi sau được lợi từ hoạt động gián điệp và những thông tin được công bố rộng rãi. Chỉ riêng việc chứng minh được tính khả thi của một sáng chế cũng có thể khuyến khích những nước khác độc lập phát triển nó; và nỗi sợ bị tụt hậu có thể sẽ làm phát sinh các nỗ lực rượt đuổi.

Có lẽ gần với trường hợp AI hơn là những phát minh toán học không đòi hỏi việc phát triển hạ tầng vật lý mới. Thông thường những phát minh này được công bố trong các tài liệu hàn lâm, và do đó có thể được coi là sẵn có; nhưng trong một số trường hợp khi mà khám phá có vẻ như cung cấp một ưu thế chiến lược, việc công bố sẽ bị trì hoãn. Ví dụ, hai trong số các ý tưởng quan trọng nhất của mã hóa khóa công khai (public key cryptography) là giao thức trao đổi khóa Diffie-Hellman và phương thức mã hóa RSA. Những ý tưởng này được khám phá ra bởi cộng đồng hàn lâm vào các năm 1976 và 1978, nhưng sau đó đã có thông tin xác nhận rằng chúng đã được các nhà

mật mã học tại nhóm an ninh truyền thông Vương quốc Anh biết đến từ đầu những năm 1970.¹²

Các dự án phần mềm lớn có thể cung cấp những trường hợp tương đồng hơn với các dự án AI, nhưng khó đưa ra những ví dụ tiêu biểu về sự tụt hậu, vì phần mềm thông thường được triển khai qua các đợt có quy mô tăng dần và các chức năng của những hệ thống đang cạnh tranh nhau không phải lúc nào cũng trực tiếp tương đồng.

Bảng 7 Một số cuộc chạy đua công nghệ có ý nghĩa chiến lược

	Mỹ	Liên Xô	Anh	Pháp	Trung Quốc	Ấn Độ	Israel	Pakistan	Triều Tiên	Nam Phi
Bom hạt nhân	1945	1949	1952	1960	1964	1974	1979?	1998	2006	1979?
Bom nhiệt hạch	1952	1953 ¹³	1957	1968	1967	1998	?	-	-	-
Khả năng phóng vệ tinh	1958	1957	1971	1965	1970	1980	1988	-	1998? ¹⁴	- ¹⁵
Khả năng đưa con người lên vũ trụ	1961	1961	-	-	2003	-	-	-	-	-
Tên lửa đạn đạo liên lục địa (ICBM) ¹⁶	1959	1960	1968 ¹⁷	1985	1971	2012	2008	- ¹⁸	2006	- ¹⁹
Tên lửa mang nhiều đầu đạn hạt nhân (MIRV) ²⁰	1970	1975	1979	1985	2007	2014 ²¹	2008?			

Khi kết hợp những quan sát này với những trao đổi trước đây của chúng ta về tốc độ cất cánh, chúng ta có thể kết luận rằng rất ít có khả năng hai dự án có thể đủ gần nhau để đồng thời cất cánh nhanh; với quá trình cất cánh trung bình, mọi việc có thể dễ dàng xảy ra theo một trong hai cách; và đối với quá trình cất cánh chậm, khả

năng rất cao là một số dự án có thể song song trải qua quá trình này. Nhưng việc phân tích cần một bước đi xa hơn. Câu hỏi quan trọng không phải là có bao nhiêu dự án cùng cất cánh, mà là có bao nhiêu dự án xuất hiện ở phía bên kia được co cụm lại về mặt năng lực đủ để không dự án nào trong số đó có lợi thế cạnh tranh quyết định. Nếu quá trình cất cánh bắt đầu tương đối chậm và sau đó tăng tốc nhanh hơn, khoảng cách giữa các dự án cạnh tranh sẽ có xu hướng tăng lên. Quay trở lại với phép ẩn dụ về những người đua xe đạp, tình hình sẽ giống với một cặp vận động viên đang đạp xe lên dốc, người này đi sau người kia một quãng – khoảng cách giữa họ sau đó sẽ tăng lên khi người đi đầu lên tới đỉnh và bắt đầu tăng tốc xuống dốc.

Hãy xem xét kịch bản cất cánh trung bình sau đây. Giả sử một dự án phải mất một năm để tăng năng lực AI từ đường cơ sở con người lên siêu trí tuệ mạnh, và một dự án bước vào giai đoạn cất cánh sớm hơn sáu tháng so với dự án tiên tiến nhất đi sau nó. Hai dự án sẽ cùng trải qua quá trình cất cánh. Khi đó, có vẻ như không dự án nào có được lợi thế cạnh tranh quyết định. Nhưng sự việc không hẳn sẽ xảy ra như vậy. Giả sử cần khoảng chín tháng để tiến từ đường cơ sở con người đến điểm quá độ, và ba tháng nữa từ đó đến siêu trí tuệ mạnh. Dự án đi trước khi đó có được siêu trí tuệ mạnh nhanh hơn ba tháng trước khi dự án tiếp theo mới chỉ đến được điểm quá độ. Điều này cho dự án dẫn đầu một lợi thế cạnh tranh quyết định và cơ hội biến vị thế dẫn đầu của nó thành sự kiểm soát vĩnh viễn bằng cách vô hiệu hóa những dự án cạnh tranh và thiết lập một thể đơn nhất (lưu ý rằng khái niệm thể đơn nhất là một khái niệm trừu tượng: một thể đơn nhất có thể là một nền dân chủ, một chế độ độc tài, một AI thống trị duy nhất, một tập hợp chuẩn mực toàn cầu mạnh mẽ có

các điều khoản có hiệu lực về việc thực thi chính nó, hoặc thậm chí là một vị chúa tể người hành tinh khác – đặc tính của nó chỉ đơn giản là một loại hình tổ chức có thể giải quyết được tất cả những vấn đề phối hợp toàn cầu quan trọng. Nó có thể, nhưng không nhất thiết, giống với bất kỳ hình thức cai trị nào của loài người.²²⁾

Vì khả năng tăng trưởng bùng nổ là đặc biệt cao ngay sau khi vượt qua điểm quá độ, khi mà vòng lặp hồi tiếp dương mạnh của công suất tối ưu hóa nhập cuộc, một kịch bản kiểu như vậy rõ ràng khả thể, và nó tăng thêm cơ hội của việc dự án sẽ chiếm được lợi thế cạnh tranh quyết định ngay cả khi quá trình cất cánh diễn ra không thực sự nhanh.

Dự án thành công sẽ lớn tới cỡ nào?

Một số con đường dẫn đến siêu trí tuệ đòi hỏi những nguồn tài nguyên to lớn, và do đó có khả năng trở thành khu vực dự trữ của những dự án lớn có nguồn vốn dồi dào. Ví dụ, giả lập hoàn chỉnh não bộ cần có nhiều loại kiến thức chuyên môn khác biệt và rất nhiều thiết bị. Tăng cường trí tuệ sinh học và giao diện não-máy tính sẽ có hệ số mở rộng lớn: trong khi một công ty kỹ thuật sinh học nhỏ có thể sáng chế ra một hai loại thuốc, việc đạt được siêu trí tuệ trên một trong những con đường này (nếu có thể thực hiện được) có lẽ sẽ đòi hỏi nhiều sáng chế và nhiều phép thử, và do đó cần đến sự hỗ trợ của một ngành công nghiệp hoặc một chương trình quốc gia được cấp đủ vốn. Đạt được siêu trí tuệ tập thể bằng cách làm cho các tổ chức và mạng lưới hiệu quả hơn còn đòi hỏi đầu vào rộng hơn liên quan đến sự tham gia của đại bộ phận nền kinh tế thế giới.

Con đường AI khó đánh giá hơn. Có thể nó sẽ đòi hỏi một chương trình nghiên cứu rất lớn; hoặc cũng có thể chỉ được làm bởi

một nhóm nhỏ. Chúng ta cũng không thể loại trừ khả năng một tin tặc đơn độc sẽ đi theo con đường này. Xây dựng một AI hạt giống có thể đòi hỏi những kiến thức và thuật toán được phát triển qua nhiều thập kỷ bởi cộng đồng các nhà khoa học trên thế giới. Nhưng có thể ý tưởng đột phá tối quan trọng lại xuất phát từ một cá nhân đơn lẻ hoặc một nhóm nhỏ gắn kết thành công mọi thứ lại với nhau. Kịch bản này ít thực tế hơn với một vài kiến trúc AI so với những kiến trúc khác. Một hệ thống có nhiều thành phần cần được điều chỉnh để có thể làm việc hiệu quả với nhau, rồi được tải vào một cách vất vả những nội dung nhận thức được tùy chỉnh, có thể sẽ đòi hỏi một dự án lớn hơn. Nhưng nếu một AI hạt giống có thể được khởi tạo như một hệ thống đơn giản, với cấu trúc chỉ phụ thuộc vào việc đáp ứng đúng một số ít nguyên tắc cơ bản, thì thành công hoàn toàn có thể nằm trong tầm với của của một nhóm nhỏ hay một cá nhân. Xác suất của việc một đột phá cuối cùng thuộc về một nhóm dự án nhỏ sẽ tăng lên nếu tiến bộ mới nhất trước đó trong lĩnh vực đã được công bố trong các tài liệu mở hoặc có sẵn trong những phần mềm mã nguồn mở.

Chúng ta phải phân biệt câu hỏi về quy mô của dự án trực tiếp thiết kế chế tạo hệ thống và câu hỏi về quy mô của nhóm sẽ kiểm soát cách thức và thời điểm tạo ra hệ thống đó. Bom nguyên tử được tạo ra chủ yếu bởi một nhóm các nhà khoa học và kỹ sư (Dự án Manhattan vào giai đoạn cao điểm sử dụng đến 130.000 người, phần lớn trong số họ là công nhân xây dựng hay nhân viên điều hành tòa nhà.²³) Tuy nhiên các chuyên gia kỹ thuật này là do quân đội Mỹ kiểm soát, quân đội lại bị kiểm soát bởi chính phủ Mỹ, cơ quan xét cho

cùng phải chịu trách nhiệm trước các cử tri Mỹ, khi đó có số lượng chỉ bằng một phần 10 dân số trưởng thành của thế giới.²⁴

Giám sát

Căn cứ vào những tác động an ninh mang tính cực đoan của siêu trí tuệ, các chính phủ có thể sẽ tìm cách quốc hữu hóa bất cứ dự án nào trên lãnh thổ của mình mà họ nghĩ rằng sắp tiến đến quá trình cất cánh. Một cường quốc cũng có thể tìm cách lấy được các dự án nằm ở lãnh thổ các quốc gia khác nhờ gián điệp, trộm cắp, bắt cóc, hối lộ, đe dọa, xâm chiếm bằng quân sự hoặc bất cứ biện pháp sẵn có nào khác. Một cường quốc không thể lấy được một dự án của nước ngoài thay vào đó có thể phá hoại nó, đặc biệt nếu nước chủ nhà không có những biện pháp răn đe hiệu quả. Nếu cấu trúc quản trị toàn cầu đủ mạnh vào thời điểm sự đột phá bắt đầu đến gần thì có thể các dự án hứa hẹn sẽ được đặt dưới dự kiểm soát quốc tế.

Do đó, một câu hỏi quan trọng là liệu các cơ quan có thẩm quyền cấp quốc gia hay quốc tế có nhìn thấy được một sự bùng nổ trí tuệ đang đến. Hiện nay, các tổ chức tình báo có vẻ như không tìm kiếm quá gắt gao những dự án AI hứa hẹn hay những dạng thức khác của việc khuếch đại trí tuệ có tiềm năng bùng nổ.²⁵ Nếu như họ thực sự không (quá) chú ý, điều này có thể là một quan điểm chung khá phổ biến cho rằng không tồn tại viễn cảnh về bất kỳ loại siêu trí tuệ nào sắp đến. Nếu, và khi các nhà khoa học có uy tín bắt đầu tin rằng có cơ hội đáng kể để siêu trí tuệ xuất hiện, các cơ quan tình báo lớn trên thế giới có thể sẽ bắt đầu giám sát các tổ chức và cá nhân có vẻ như sẽ tham gia vào các nghiên cứu liên quan. Bất cứ dự án nào bắt đầu cho thấy tiến bộ thích hợp đều có thể nhanh chóng bị quốc hữu hóa. Nếu giới chính trị gia cấp cao bị thuyết phục bởi tính nghiêm trọng

của rủi ro, những nỗ lực dân sự trong các khu vực nhạy cảm có thể sẽ bị chình đốn hoặc cấm đoán.

Vậy việc giám sát này khó đến đâu? Nhiệm vụ sẽ dễ dàng hơn nếu mục tiêu chỉ là theo dõi dự án dẫn đầu. Trong trường hợp này, việc giám sát có thể chỉ cần tập trung vào một số dự án được cấp nguồn tài nguyên tốt nhất. Nếu thay vào đó, mục tiêu là ngăn chặn mọi công việc (ít nhất là bên ngoài những thể chế được cấp phép đặc biệt), việc giám sát có thể phải toàn diện hơn, vì nhiều dự án nhỏ và cá nhân có thể ít nhất có được sự tiến bộ nào đó.

Việc giám sát các dự án đòi hỏi lượng vốn vật lý lớn sẽ dễ dàng hơn, như trường hợp của một dự án giả lập hoàn chỉnh não bộ. Ngược lại, việc nghiên cứu trí tuệ nhân tạo chỉ cần có một máy tính cá nhân, và do đó sẽ khó giám sát hơn. Một số công trình lý thuyết có thể được thực hiện mà chỉ cần giấy bút. Kể cả như vậy, sẽ không quá khó khăn để xác định những cá nhân có năng lực nhất có quan tâm nghiêm túc và lâu dài tới việc nghiên cứu trí tuệ nhân tạo. Những cá nhân này thường để lại những dấu vết có thể nhìn thấy được. Họ có thể có các bài báo khoa học được công bố, phát biểu tại các hội nghị, đăng bài trên các diễn đàn Internet hoặc tốt nghiệp các khoa về khoa học máy tính hàng đầu trên thế giới. Họ cũng có thể có liên lạc với các nhà nghiên cứu AI khác. Việc này cho phép nhận dạng họ bằng cách lập biểu đồ xã hội.

Những dự án được thiết kế bí mật ngay từ đầu có thể khó phát hiện hơn. Một dự án phát triển phần mềm tầm thường cũng có thể được dùng làm bình phong.²⁶ Chỉ có phân tích kỹ lưỡng mã phần mềm đang được tạo ra mới làm hé lộ bản chất thực sự của điều mà dự án đang cố gắng thực hiện. Những phân tích như vậy thường đòi

hỏi rất nhiều nhân lực (có kỹ năng cao), nên chỉ một số ít dự án có thể được rà soát ở cấp độ này. Nhiệm vụ có thể trở nên dễ dàng hơn nhiều nếu công nghệ phát hiện nói dối hiệu quả được phát triển và có thể sử dụng thường xuyên trong những loại hình giám sát này.²⁷

Một lý do khác khiến các quốc gia không nắm bắt được những phát triển mang tính tiền thân là sự khó khăn cố hữu của việc dự báo một số kiểu đột phá. Điều này liên quan đến việc nghiên cứu AI nhiều hơn là phát triển giả lập hoàn chỉnh não bộ, vì đối với loại hình thứ hai, những đột phá quan trọng thường có thể được mở đường bằng một chuỗi rõ ràng những tiến bộ chắc chắn.

Cũng có thể các tổ chức tình báo và cơ quan chính phủ khác ở một mức độ nào đó khá vụng về và cứng nhắc nên không hiểu được ý nghĩa của một số diễn biến có thể rất rõ ràng đối với các nhóm khác bên ngoài. Những rào cản đối với hiểu biết chính thức về một sự bùng nổ trí tuệ tiềm tàng có thể rất lớn. Chẳng hạn, ta có thể hình dung được việc một chủ đề có dính dáng đến mâu thuẫn về tôn giáo hay chính trị, khiến nó trở thành điều cấm kỵ ở một số quốc gia. Chủ đề đó có thể sẽ trở nên dính líu đến một số nhân vật tai tiếng, lừa đảo hoặc một sự thổi phồng nói chung nào đó, và vì thế bị các nhà khoa học có uy tín và những nhân vật nổi tiếng khác xa lánh. (Như chúng ta thấy trong Chương 1, việc tương tự đã xảy ra hai lần: hãy nhớ lại về hai “mùa đông AI”). Các tập đoàn công nghiệp có thể vận động hành lang để phòng ngừa việc những lĩnh vực kinh doanh có lợi nhuận bị bôi xấu; cộng đồng hàn lâm có thể xiết chặt đội ngũ để loại bỏ những người nêu lên quan ngại về các hệ quả dài hạn của ngành khoa học đang được nghiên cứu.²⁸

Về mặt hậu quả, sự thất bại hoàn toàn về trí tuệ là không thể tránh khỏi. Thất bại này đặc biệt có khả năng xảy ra nếu các đột phá diễn ra trong tương lai gần, trước khi vấn đề trở thành tâm điểm chú ý. Và thậm chí nếu các tổ chức tình báo hiểu được vấn đề thì các nhà lãnh đạo chính trị có thể cũng sẽ không lắng nghe hay hành động theo những lời tư vấn. Quá trình khởi động Dự án Manhattan cần đến nỗ lực phi thường của một vài nhà vật lý nhìn xa trông rộng, đặc biệt là Mark Oliphant và Leó Szilárd. Szilárd đã thuyết phục Eugene Wigner để ông này thuyết phục Albert Einstein ký tên vào bức thư mà sau đó được dùng để thuyết phục Tổng thống Franklin D. Roosevelt quan tâm đến vấn đề này. Ngay cả sau khi dự án đã đạt đến quy mô hoàn chỉnh, Roosevelt cũng vẫn còn hoài nghi về tính khả thi và ý nghĩa của nó, cũng giống như người kế nhiệm ông là Tổng thống Harry Truman.

Dù sao chăng nữa, một nhóm nhỏ các nhà hoạt động có thể sẽ khó tác động hơn vào kết quả của một sự bùng nổ trí tuệ nếu những “gã khổng lồ” (như nhà nước) đóng vai trò chủ động. Vì thế, các cơ hội để các cá nhân giảm thiểu mức độ rủi ro toàn diện mang tính sống còn phát sinh từ sự bùng nổ trí tuệ tiềm tàng là lớn nhất ở các kịch bản mà trong đó những gã khổng lồ tương đối ít chú ý tới vấn đề, hoặc trong đó, những nỗ lực ban đầu của các nhà hoạt động tạo được khác biệt quan trọng lên việc những người chơi lớn tham gia vào sân chơi là có hay không, khi nào và với thái độ nào. Vì vậy, các nhà hoạt động tìm kiếm tác động kỳ vọng tối đa có thể muốn tập trung hầu hết việc lập kế hoạch của họ vào những kịch bản có tác động đòn bẩy lớn, ngay cả khi họ tin rằng kịch bản mà trong đó những gã khổng lồ quyết định tất cả cũng có thể xảy ra.

Hợp tác quốc tế

Việc hợp tác quốc tế có khả năng xảy ra hơn nếu các cấu trúc quản trị toàn cầu nói chung trở nên mạnh mẽ hơn. Việc hợp tác này còn có khả năng xảy ra cao hơn nữa nếu ý nghĩa của một sự bùng nổ trí tuệ được biết trước và nếu có thể giám sát hiệu quả tất cả các dự án nghiêm túc. Tuy nhiên, ngay cả khi việc giám sát là bất khả thi thì sự hợp tác quốc tế cũng vẫn có thể diễn ra. Nhiều quốc gia có thể liên kết với nhau để hỗ trợ một dự án chung. Nếu dự án chung này được cấp đủ nguồn tài nguyên, nó có thể có cơ hội tốt để trở thành dự án đầu tiên đạt được mục đích, đặc biệt nếu bất cứ dự án đối địch nào phải có quy mô nhỏ và bí mật để tránh bị phát hiện.

Trước đây đã từng có một số dự án hợp tác khoa học đa quốc gia quy mô lớn thành công, như Trạm Vũ trụ Quốc tế, Dự án bản đồ gen người và Máy gia tốc hạt Hadron lớn.²⁹ Tuy nhiên, động cơ chính để hợp tác trong những trường hợp này là chia sẻ chi phí. (Với Trạm Vũ trụ quốc tế, bản thân việc nuôi dưỡng tinh thần hợp tác giữa Nga và Mỹ cũng là một mục tiêu quan trọng.³⁰) Sự hợp tác như vậy sẽ khó đạt được hơn trong một dự án có nhiều ý nghĩa bảo mật. Một quốc gia tin tưởng rằng mình có thể đơn phương đạt được sự đột phá sẽ muốn đơn thương độc mã thay vì dành một phần công sức để tham gia vào một dự án chung. Một quốc gia có thể không tham gia một dự án hợp tác quốc tế vì sợ rằng các bên tham gia khác “hút” hết những thông tin quan trọng được chung tay tạo ra và sử dụng chúng để tăng tốc một dự án quốc gia bí mật.

Một dự án quốc tế, vì thế, cần phải vượt qua những thách thức lớn về an ninh và một niềm tin tương đối lớn có thể cũng là cần thiết để khởi động nó, một niềm tin cần nhiều thời gian để xây dựng. Lưu

ý rằng thậm chí sau sự tan vỡ trong quan hệ Nga-Mỹ sau khi Gorbachev lên cầm quyền, các nỗ lực cắt giảm vũ khí – được cả hai cường quốc quan tâm – đã có một sự khởi đầu không ổn định. Gorbachev tìm cách cắt giảm nhanh chóng vũ khí hạt nhân, nhưng các cuộc đàm phán đã mắc kẹt ở vấn đề Sáng kiến phòng thủ chiến lược của Reagan (“Chiến tranh giữa các vì sao”), vấn đề mà điện Kremlin cực lực phản đối. Tại hội nghị thượng đỉnh Reykjavik năm 1986, Reagan đề nghị Mỹ chia sẻ với Liên Xô công nghệ sẽ được phát triển theo Sáng kiến phòng thủ chiến lược, sao cho cả hai quốc gia cùng được bảo vệ chống lại những vụ phóng tên lửa do sự cố và những quốc gia nhỏ hơn có thể đã phát triển được vũ khí hạt nhân. Mặc dù vậy, Gorbachev vẫn không được thuyết phục bởi đề xuất hai bên cùng có lợi này. Ông nhìn nhận vụ việc này như một âm mưu và từ chối công nhận ý tưởng rằng người Mỹ có thể chia sẻ thành quả từ những nghiên cứu quân sự tiên tiến nhất của họ khi mà họ thậm chí còn không mong muốn chia sẻ với Liên Xô công nghệ vắt sữa bò.³¹

Bất kể Reagan trên thực tế có chân thành hay không trong đề nghị của mình về sự hợp tác giữa các siêu cường, việc mất lòng tin khiến cho đề xuất này trở nên vô nghĩa.

Sự hợp tác sẽ dễ đạt được hơn giữa các đồng minh nhưng cũng không phải tự nhiên mà có. Khi Liên Xô và Mỹ kết thành đồng minh chống lại Đức trong Thế Chiến II, Mỹ đã giấu Liên Xô dự án bom nguyên tử của mình, nhưng có cộng tác với Anh và Canada trong Dự án Manhattan.³² Tương tự, Anh đã giấu Liên Xô thành công của mình trong việc giải mã Enigma của Đức nhưng lại chia sẻ nó với Mỹ.³³ Điều này gợi ý rằng để đạt được hợp tác quốc tế trong một

công nghệ nào đó có tầm quan trọng đặc biệt đối với an ninh quốc gia, có lẽ trước tiên phải cần đến một mối quan hệ gần gũi và đáng tin cậy.

Chúng ta sẽ quay trở lại với tính hấp dẫn và khả thi của hợp tác quốc tế trong việc phát triển các công nghệ khuếch đại trí tuệ trong Chương 14.

Từ lợi thế cạnh tranh quyết định đến thể đơn nhất

Liệu một dự án có được lợi thế cạnh tranh quyết định có chọn sử dụng ưu thế này để hình thành một thể đơn nhất hay không?

Hãy xem xét một tình huống lịch sử có phần tương tự. Mỹ phát triển được vũ khí hạt nhân năm 1945 và đó đã từng là vũ khí hạt nhân duy nhất cho tới khi Liên Xô phát triển được bom nguyên tử năm 1949. Trong giai đoạn này – và một khoảng thời gian sau đó – Mỹ có thể có, hoặc ở vị thế có thể đạt được, một ưu thế quân sự quyết định.

Khi đó, về lý thuyết, Mỹ có thể sử dụng vị thế độc quyền về nguyên tử để tạo ra thể đơn nhất. Họ có thể làm việc này bằng cách dốc hết sức thực hiện một nỗ lực để xây dựng kho vũ khí nguyên tử, và sau đó đe dọa (và tiến hành nếu thấy cần thiết) một cuộc tấn công hạt nhân đầu tiên để phá hoại năng lực công nghiệp của bất cứ chương trình hạt nhân mới chớm nở nào tại Liên Xô hay bất cứ quốc gia nào khác mong muốn phát triển năng lực hạt nhân.

Quá trình hành động “lành tính” hơn và có thể có tác dụng sẽ là sử dụng kho vũ khí nguyên tử của mình như một lợi điểm thương lượng để đàm phán tạo ra một chính quyền quốc tế mạnh – một cơ chế Liên Hợp Quốc phi phủ quyết với vị thế độc quyền về nguyên tử và nghĩa vụ thực hiện mọi hành động cần thiết để ngăn chặn bất cứ quốc gia nào tự phát triển vũ khí hạt nhân. Cả hai cách tiếp cận này

khi đó đều đã được đề xuất. Cách tiếp cận cứng rắn của việc tung ra hay đe dọa tấn công trước được một số trí thức nổi tiếng khi đó ủng hộ, chẳng hạn như Bertrand Russell (người từ lâu đã rất tích cực trong các phong trào phản chiến và sau này đã dành nhiều thập kỷ tranh đấu chống vũ khí nguyên tử) cùng John von Neuman (người đồng phát minh lý thuyết trò chơi và một trong những kiến trúc sư của chiến lược nguyên tử Mỹ).³⁴ Có lẽ một dấu hiệu của sự tiến bộ về văn minh chính là việc ý tưởng đe dọa đánh chặn hạt nhân (first strike) ngày nay xem ra đang dần được coi là ngớ ngếch hoặc phản cảm về mặt đạo đức.

Một phiên bản của cách tiếp cận “lành tính” này đã được Mỹ thử nghiệm năm 1946 trong kế hoạch Baruch. Đề xuất của kế hoạch liên quan đến việc Mỹ từ bỏ vị thế độc quyền tạm thời của mình. Việc khai thác urani và thori cùng công nghệ nguyên tử sẽ được đặt dưới sự kiểm soát của một tổ chức quốc tế hoạt động dưới sự bảo hộ của Liên Hợp Quốc. Đề xuất kêu gọi thành viên thường trực của Hội đồng Bảo an Liên Hợp Quốc từ bỏ quyền phủ quyết trong các vấn đề liên quan đến vũ khí hạt nhân để ngăn chặn bất cứ cường quốc nào vi phạm hiệp ước phủ quyết việc áp dụng các biện pháp khắc phục.³⁵ Khi nhận thấy rằng Liên Xô và các đồng minh có thể dễ dàng bị thua số phiếu trong cả Hội đồng Bảo an và Đại Hội đồng, Stalin đã từ chối đề xuất này. Một không khí băng giá của việc nghi ngờ lẫn nhau đã bao trùm lên các mối quan hệ giữa những cựu đồng minh thời chiến, sự mất tin tưởng này sau đó đã “cô đặc” lại thành Chiến tranh Lạnh. Và như đã được dự báo rộng rãi từ trước, kéo theo đó là một cuộc chạy đua vũ trang tốn kém và vô cùng nguy hiểm.

Nhiều yếu tố có thể ngăn trở việc một tổ chức của con người có ưu thế chiến lược quyết định tạo ra thể đơn nhất. Các yếu tố này bao gồm các hàm thỏa dụng hạn chế (bounded) hoặc không gộp (non-aggregative), các quy tắc ra quyết định phi tối đa hóa, sự hỗn loạn và bất định, các vấn đề phối hợp và những chi phí khác nhau liên quan đến việc tiếp quản. Nhưng nếu đó lại là một tổ chức không phải của con người mà là một tác tử trí tuệ nhân tạo có được ưu thế chiến lược quyết định? Liệu các yếu tố được nhắc đến ở trên có tác dụng tương đương trong việc ngăn cản AI tìm cách chiếm quyền? Hãy xem qua một danh sách các yếu tố và cân nhắc về cách mà chúng có thể áp dụng trong trường hợp này.

Các cá nhân và tổ chức của con người thường có ưu tiên đối với những nguồn tài nguyên không được thể hiện rõ ràng bởi một “hàm thỏa dụng gộp không hạn chế” (unbounded aggregative utility function). Một người thông thường không đánh cược toàn bộ vốn liếng của mình cho một cơ hội nhân đôi số tiền có tỷ lệ thành công 50-50. Một quốc gia thường không chấp nhận rủi ro mất toàn bộ lãnh thổ của mình để đổi lại cơ hội mở rộng gấp 10 lần chỉ thành công 10%. Đối với các chính phủ và cá nhân, quy luật hiệu suất giảm dần đúng với hầu hết các nguồn tài nguyên, nhưng điều này không đúng với AI. (Chúng ta sẽ quay trở lại với vấn đề tạo động lực cho AI trong những chương sau). Một AI do đó sẽ dễ có khả năng theo đuổi một quỹ đạo hành động rủi ro hơn nhưng có thể cho nó cơ hội kiểm soát thế giới.

Con người và các tổ chức do con người điều hành có thể vận hành với các quy trình ra quyết định không nhằm mục đích tối đa hóa độ thỏa dụng mong đợi. Ví dụ, họ có thể cho phép một mức độ ngại

rủi ro cơ bản, hay các quy tắc ra quyết định “thỏa mãn vừa đủ” (satisficing) tập trung vào đạt được các ngưỡng đầy đủ (adequacy threshold), hoặc những rào cản phụ mang tính “nghĩa vụ” ngăn cấm một số loại hành động bất kể kết quả của chúng được mong muốn đến đâu. Những người ra quyết định xem ra thường hành động theo một bản ngã (identity) hoặc một vai trò xã hội nào đó thay vì tìm cách tối đa hóa việc đạt được một mục tiêu cụ thể. Nhắc lại, điều này không cần áp dụng cho các tác tử nhân tạo.

Hàm thỏa dụng hạn chế, tính ngại rủi ro và các quy tắc quyết định phi tối đa hóa có thể kết hợp một cách cộng lực với sự hỗn loạn và bất định chiến lược. Các cuộc cách mạng, ngay cả khi thành công trong việc lật đổ trật tự hiện tại, cũng thường thất bại trong việc đem lại những kết quả mà những người khởi xướng đã hứa hẹn. Nếu hành động đã dự liệu là không thể đảo ngược, phá vỡ tiêu chuẩn hay không có tiền lệ thì thông thường nó sẽ có xu hướng không được con người thực hiện. Một siêu trí tuệ có thể nhìn nhận tình hình rõ ràng hơn, do đó ít gặp phải sự hỗn loạn và bất định chiến lược hơn về kết quả, nếu như nó tìm cách sử dụng ưu thế chiến lược quyết định hiển nhiên của mình để tạo lập vị thế thống trị.

Một yếu tố quan trọng khác có thể cản trở các nhóm khai thác một ưu thế chiến lược quyết định tiềm tàng là vấn đề phối hợp nội bộ. Các thành viên của một nhóm đang âm mưu chiếm quyền có thể lo lắng không chỉ về việc bị thâm nhập từ bên ngoài mà còn bị lật đổ bởi một liên minh nhỏ nào đó của những người trong cuộc. Nếu một nhóm được hình thành từ 100 người, và 60 người (đa số) có thể chiếm được quyền lực và tước quyền lợi của những người không đồng mưu, thì liệu có điều gì ngăn chặn được việc nhóm nhỏ 35

người trong số 60 người đó trước đoạt quyền lợi của 25 người còn lại, và sau đó là việc nhóm nhỏ 20 người lại đoạt trước quyền lợi của 15 người kia? Từng người trong nhóm 100 người ban đầu có thể có lý do phù hợp để ủng hộ một số tiêu chuẩn chính thức nhất định, nhằm ngăn chặn sự đổ vỡ tổng thể có thể xuất phát từ bất cứ nỗ lực thay đổi khế ước xã hội nào thông qua việc thu tóm quyền lực thô bạo. Vấn đề phối hợp nội bộ này không áp dụng với một hệ thống AI chỉ bao gồm một tác tử hợp nhất duy nhất.³⁶

Cuối cùng là vấn đề chi phí. Ngay cả nếu Mỹ có thể sử dụng sự độc quyền về nguyên tử của mình để thiết lập một thể đơn nhất, thì đất nước này hẳn cũng không làm được điều đó mà không phát sinh những chi phí đáng kể. Trong trường hợp thỏa thuận được đàm phán để đặt vũ khí hạt nhân dưới dự kiểm soát của Liên Hợp Quốc đã cải cách và củng cố, những chi phí này có thể tương đối thấp; nhưng những cái giá khác – đạo đức, kinh tế, chính trị và con người – của việc thực sự tìm cách chinh phục thế giới bằng chiến tranh hạt nhân sẽ lớn tới mức không thể tưởng tượng nổi, thậm chí cả trong giai đoạn độc quyền hạt nhân. Tuy nhiên, khi có đủ mức độ vượt trội về công nghệ, những cái giá này có thể sẽ nhỏ hơn rất nhiều. Ví dụ, hãy xem xét một kịch bản mà trong đó, một quốc gia có sự vượt trội công nghệ lớn tới mức có thể trước vũ khí của tất cả các quốc gia khác một cách an toàn chỉ bằng một lần bấm nút, không khiến bất kỳ ai thiệt mạng hay bị thương, và hầu như không gây tác hại với môi trường hoặc cơ sở hạ tầng. Với sự vượt trội công nghệ thần kỳ này, việc đánh chặn hạt nhân sẽ hấp dẫn hơn rất nhiều. Hoặc, hãy xem xét một mức độ vượt trội công nghệ thậm chí còn lớn hơn nữa, có thể cho phép người đi đầu buộc các quốc gia khác tự nguyện hạ vũ

khí, không phải bằng cách đe dọa tiêu diệt họ mà chỉ đơn giản là thuyết phục phần lớn người dân của họ bằng các chiến dịch quảng cáo và tuyên truyền được thiết kế siêu hiệu quả về một nền thống nhất toàn cầu. Nếu điều đó được thực hiện với ý định mang lại lợi ích cho mọi người, ví dụ như bằng cách thay thế sự đối đầu giữa các quốc gia và chạy đua vũ trang bằng một chính phủ thế giới chung công bằng, đại diện và hiệu quả, thì cũng chưa rõ liệu có hay không một sự phản đối đạo đức có tính thuyết phục đối với việc biến một ưu thế chiến lược tạm thời thành một thể đơn nhất vĩnh viễn.

Những cân nhắc khác nhau, do đó, đều chỉ ra một khả năng ngày càng tăng của việc một cường quốc tương lai có siêu trí tuệ chiếm được ưu thế chiến lược đủ lớn sẽ thực sự sử dụng ưu thế này để tạo ra thể đơn nhất. Mức độ đáng mong muốn của một kết quả như vậy tất nhiên còn phụ thuộc vào bản chất của thể đơn nhất có thể được tạo ra, và việc tương lai của sự sống thông minh sẽ như thế nào trong những kịch bản thay thế đa cực. Chúng ta sẽ quay lại với những câu hỏi này trong các chương sau, nhưng trước hết, hãy xem xét kỹ lưỡng hơn lý do và cách thức để một siêu trí tuệ trở nên mạnh mẽ và hiệu quả trong việc đạt được các kết quả trên thế giới.

Chương 6

Siêu năng lực nhận thức

Gia sử một phần mềm siêu trí tuệ số ra đời, và vì lý do nào đó muốn chiếm quyền kiểm soát thế giới: liệu nó có thể làm được hay không? Trong chương này, chúng ta sẽ xem xét một vài năng lực mà siêu trí tuệ có thể phát triển và những gì mà các năng lực này cho phép nó thực hiện. Chúng tôi sẽ lược ra một kịch bản chiếm quyền minh họa cách thức mà một tác tử siêu trí tuệ, chỉ bắt đầu như một phần mềm, có thể tự đặt mình vào thế đơn nhất. Chúng tôi cũng sẽ đưa ra vài nhận xét về mối quan hệ giữa năng lực đối với tự nhiên và năng lực đối với những tác tử khác.

Lý do cơ bản cho sự thống trị của loài người trên Trái đất là bộ não của chúng ta có một số khả năng mở rộng hơn đôi chút so với những loài động vật khác.¹ Trí tuệ mạnh hơn cho phép chúng ta truyền tải văn hóa hiệu quả hơn, kết quả là kiến thức và công nghệ được tích lũy từ thế hệ này sang thế hệ khác. Cho tới nay, chúng ta đã tích lũy đủ nội dung để có thể bay vào vũ trụ, chế tạo bom nhiệt hạch, áp dụng kỹ thuật di truyền, chế tạo máy tính, xây dựng các khu nhà xưởng, làm thuốc trừ sâu, khơi dậy các phong trào hòa bình quốc tế và tạo ra toàn bộ trang thiết bị của nền văn minh hiện đại. Các nhà địa chất đã bắt đầu gọi thời kỳ hiện tại là Anthropocene (thế Nhân Sinh) để ghi nhận những dấu ấn sinh học, trầm tích và địa hóa

của các hoạt động của con người.² Theo một ước tính, con người đã chiếm 24% tổng lượng sản xuất cơ bản thuần của hệ sinh thái trên hành tinh.³ Mặc dù vậy, chúng ta vẫn còn xa mới đạt đến giới hạn vật lý của công nghệ.

Những quan sát này khả thi hóa việc bất cứ kiểu thực thể nào phát triển tới mức trí tuệ vượt xa con người đều có tiềm năng trở nên vô cùng mạnh mẽ. Những thực thể này có thể tích lũy nội dung nhanh hơn chúng ta rất nhiều và sáng chế những công nghệ mới trong những khoảng thời gian ngắn hơn đáng kể. Ngoài ra, chúng còn có thể sử dụng trí tuệ của mình để lập chiến lược hiệu quả hơn con người.

Hãy cùng nhau xem xét một số năng lực mà một siêu trí tuệ có thể có và nó có thể sử dụng chúng như thế nào.

Các chức năng và siêu năng lực

Điều quan trọng là không nên nhân hình hóa siêu trí tuệ khi nghĩ tới những tác động tiềm tàng của nó. Các khung tham chiếu nhân hình khuyến khích những kỳ vọng không có cơ sở về quỹ đạo tăng trưởng của một AI hạt giống, cũng như về tâm lý, động cơ và năng lực của một siêu trí tuệ hoàn thiện

Chẳng hạn, chúng ta rất hay giả định rằng một cỗ máy siêu trí tuệ có thể giống một người rất thông minh nhưng lập dị; và tưởng tượng rằng AI này có trí tuệ sách vở nhưng thiếu kiến thức xã hội, hoặc nó suy nghĩ logic nhưng không sáng tạo và không có trực giác. Ý tưởng này có lẽ xuất phát từ quan sát sau: chúng ta nhìn những chiếc máy tính thời hiện đại và thấy rằng chúng giỏi tính toán, ghi nhớ sự kiện và làm theo chính xác các hướng dẫn trong khi lãng quên đi hoàn cảnh và ẩn ý, cũng như các tiêu chuẩn, cảm xúc và vấn đề chính trị.

Sự liên hệ này được củng cố khi chúng ta quan sát thấy những người giỏi máy tính thường có thiên hướng lập dị. Như vậy rất tự nhiên khi giả định rằng trí tuệ máy tính tiên tiến hơn sẽ có những thuộc tính tương tự, nhưng ở mức độ cao hơn.

Sự tự nghiệm này có thể vẫn đúng trong những giai đoạn phát triển ban đầu của một AI hạt giống. (Không có bất cứ lý do nào để giả định rằng điều này sẽ áp dụng được cho các bản giả lập hoặc những người được tăng cường nhận thức). Trong giai đoạn chưa trưởng thành, thứ mà sau này sẽ trở thành một AI siêu trí tuệ có thể vẫn thiếu nhiều kỹ năng và tài năng mà con người tự nhiên có được; quy luật về điểm mạnh và điểm yếu của AI hạt giống này có thể tương đồng một cách mờ nhạt với một thiên tài lập dị. Đặc tính quan trọng nhất của một AI hạt giống, ngoài việc có thể dễ dàng cải thiện (có sức chống đối thấp), là rất giỏi sử dụng công suất tối ưu hóa để khuếch đại trí tuệ của hệ thống: một kỹ năng được coi là có liên quan mật thiết đến việc giỏi tính toán, lập trình, thiết kế, nghiên cứu khoa học máy tính và những mục đích “lập dị” khác. Tuy nhiên, ngay cả khi một AI hạt giống có một hồ sơ năng lực như vậy trong một giai đoạn phát triển nhất định, điều này cũng không đồng nghĩa với việc nó sẽ tăng trưởng để trở thành một siêu trí tuệ hoàn chỉnh với những hạn chế tương tự. Hãy nhớ lại sự khác biệt giữa tầm tiếp cận trực tiếp và gián tiếp. Với kỹ năng khuếch đại trí tuệ thích hợp, mọi năng lực trí tuệ đều nằm trong tầm tiếp cận gián tiếp: hệ thống có thể phát triển những module nhận thức và những kỹ năng mới nếu cần thiết – bao gồm sự đồng cảm, sự nhạy cảm chính trị, và bất cứ năng lực nào khác thường được mong muốn một cách rập khuôn trong các nhân cách giống máy tính.

Thậm chí nếu công nhận rằng một siêu trí tuệ có thể có tất cả những kỹ năng và tài năng thường thấy ở con người, cùng với những tài năng khác loài người không có, thì xu hướng nhân hình hóa vẫn có thể khiến chúng ta đánh giá thấp mức độ vượt trội của siêu trí tuệ máy so với hiệu năng cấp độ con người. Eliezer Yudkowsky, như đã bàn đến ở một chương phía trên, đã đặc biệt nhấn mạnh quan điểm lên án hình thức tư duy sai lầm này: những khái niệm trực giác của chúng ta về “thông minh” và “ngốc nghếch” được cất lọc từ trải nghiệm về sự khác biệt trên một “phổ” tư duy, nhưng những khác biệt về năng lực tư duy trong phổ này là không đáng kể so với những sự khác biệt giữa trí tuệ con người và siêu trí tuệ.⁴

Chương 3 đã xem xét một số nguồn tiềm năng tạo ra ưu thế của trí tuệ máy. Những tiến bộ này đủ lớn để gợi ý rằng có thể sẽ chính xác hơn nếu nghĩ AI này thông minh giống như người bình thường so với giun hay bọ, thay vì nghĩ rằng nó thông minh giống như thiên tài so với người thường.

Thật tiện lợi nếu có thể lượng hóa mức độ nhận thức của một hệ thống nhận thức bất kỳ bằng một số chỉ số quen thuộc như điểm IQ, hoặc một phiên bản nào đó của hệ thống đánh giá Elo dùng để đo lường khả năng tương đối của các đấu thủ trong những môn thi đấu hai người như cờ vua. Tuy nhiên, những chỉ số này không hữu dụng trong bối cảnh trí tuệ nhân tạo tổng thể cấp độ siêu nhân. Chúng ta không quan tâm đến xác suất một trí tuệ nhân tạo thắng một ván cờ vua cao đến mức nào. Điểm IQ thì chỉ có ý nghĩa chừng nào chúng ta có được ý tưởng về cách mà điểm số này tương quan với những kết quả liên quan đến thực tế.⁵ Ví dụ, chúng ta có dữ liệu cho thấy

những người có IQ 130 có khả năng vượt trội hơn ở trường học so với những người có IQ 90, và làm tốt những công việc đòi hỏi khả năng nhận thức cao. Nhưng giả sử chúng ta, bằng cách nào đó, có thể xác định được rằng một AI tương lai nhất định sẽ có IQ bằng 6.455: vậy thì sao? Chúng ta có thể không biết AI này làm được những gì. Chúng ta thậm chí còn không thể biết được AI này có trí tuệ tổng quát như của một người lớn bình thường hay không – có lẽ thay vào đó, AI này chỉ có một mớ thuật toán đặc thù mục tiêu, cho phép nó trả lời các câu hỏi kiểm tra trí tuệ điển hình với hiệu năng của siêu nhân, và ngoài ra không còn gì khác.

Hiện tại, con người đã thực hiện một vài nỗ lực nhằm phát triển các tham số đo lường năng lực nhận thức có thể áp dụng được với nhiều loại hệ thống xử lý thông tin, bao gồm cả trí tuệ nhân tạo.⁶

Những nỗ lực theo hướng này, nếu vượt qua được những khó khăn khác nhau về kỹ thuật, có thể trở nên khá hữu ích cho một số mục tiêu khoa học, kể cả phát triển AI. Tuy nhiên, với mục tiêu của nghiên cứu hiện tại, tính hữu ích của nó có thể bị hạn chế vì chúng ta vẫn chưa biết một mức điểm số hiệu năng nhất định của siêu trí tuệ có ý nghĩa gì đối với năng lực thực sự trong việc đạt được một kết quả quan trọng về mặt thực tế trên thế giới.

Do đó, sẽ phù hợp hơn với mục tiêu của chúng ta khi liệt kê một số tác vụ quan trọng về mặt chiến lược, sau đó xác định đặc điểm của các hệ thống nhận thức giả thuyết xem liệu chúng có bất kỳ kỹ năng nào cần để thực hiện thành công những tác vụ này hay không (xem Bảng 8). Chúng ta sẽ nói rằng một hệ thống đủ vượt trội trong bất kỳ tác vụ nào trong bảng này sẽ sở hữu một siêu năng lực tương ứng.

Một siêu trí tuệ hoàn thiện sẽ vượt trội hơn rất nhiều trong tất cả các tác vụ đó, và vì vậy sẽ có đủ cả sáu siêu năng lực. Chúng ta vẫn chưa biết rõ về khả năng xuất hiện có ý nghĩa thực tế của một trí tuệ bị giới hạn về lĩnh vực sở hữu một số siêu năng lực, nhưng chưa có được toàn bộ số đó, trong một thời gian đáng kể. Tạo ra một chiếc máy với bất cứ một siêu năng lực nào trong số đó có lẽ là vấn đề AI toàn diện. Mặc dù vậy, ta vẫn có thể hình dung được rằng một siêu trí tuệ tập thể nào đó, với số lượng đủ lớn các trí óc sinh học hay điện tử cấp độ con người, sẽ có (chẳng hạn như) siêu năng lực năng suất kinh tế nhưng không có siêu năng lực lập chiến lược. Tương tự, có thể hình dung một AI chuyên về kỹ thuật có thể được tạo ra với siêu năng lực nghiên cứu công nghệ, nhưng hoàn toàn không có kỹ năng trong những lĩnh vực khác. Điều này sẽ hợp lý hơn nếu có một lĩnh vực công nghệ cụ thể nào đó, sao cho mức độ điều luyện trong đó đủ để tạo ra một công nghệ tổng quát vượt trội áp đảo. Ví dụ, ta có thể hình dung ra một AI chuyên biệt thông thạo việc mô phỏng các hệ thống phân tử và phát minh ra những thiết kế phân tử nano có thể hiện thực hóa được nhiều năng lực quan trọng (như các máy tính hay hệ thống vũ khí với các đặc tính hiệu năng của tương lai) mới chỉ được người dùng mô tả ở mức độ trừu tượng tương đối cao.⁷ Một AI như vậy có thể có khả năng chế tạo ra một bản thiết kế chi tiết để vươn lên từ công nghệ hiện tại (như công nghệ sinh học hoặc công nghệ protein), nhằm tiến tới những năng lực tạo dựng cần thiết cho quá trình sản xuất công suất cao với độ chính xác ở mức nguyên tử, cho phép sản xuất các cấu trúc cơ học nano đa dạng hơn nhiều với chi phí thấp.⁸ Tuy nhiên, cũng có thể xảy ra khả năng một AI kỹ thuật không thể thực sự sở hữu được siêu năng lực nghiên cứu công

nghệ, và cũng không có những kỹ năng cao cấp trong các lĩnh vực ngoài công nghệ, gồm nhiều lĩnh vực trí thức sâu rộng có thể cần đến để hiểu được cách diễn dịch những yêu cầu của người dùng, cách lập mô hình hành vi của một thiết kế trong các ứng dụng thế giới thực, cách xử lý lỗi lập trình và hỏng hóc bất ngờ, cách mua sắm vật tư và đầu vào cần cho xây dựng cùng nhiều vấn đề khác.⁹

Thao túng xã hội	Lập mô hình xã hội và tâm lý, thao túng, hùng biện thuyết phục	• Tận dụng nguồn tài nguyên bên ngoài bằng cách sử dụng sự hỗ trợ của con người • Cho phép một AI bị "đóng hộp" thuyết phục người canh gác thả nó ra ngoài • Thuyết phục các nhà nước và tổ chức ứng dụng một quỹ đạo hành động nào đó
Hack	Tìm kiếm và khai thác những kẽ hở an ninh trong các hệ thống máy tính	• AI có thể trưng dụng nguồn tài nguyên tính toán trên Internet • Một AI "đóng hộp" có thể khai thác các lỗ hổng an ninh để thoát khỏi những giới hạn về điều khiển • Lấy trộm tài nguyên tài chính • Xâm chiếm hạ tầng, thao túng robot quân sự...
Nghiên cứu công nghệ	Thiết kế và lập mô hình các công nghệ tiên tiến (như công nghệ sinh học, công nghệ nano) và những con đường phát triển	• Tạo ra lực lượng quân đội mạnh mẽ • Tạo ra hệ thống giám sát • Tự động thuộc địa hóa không gian
Năng suất kinh tế	Những kỹ năng khác nhau cho phép thực hiện các công việc tri thức có năng suất kinh tế cao	• Tạo ra của cải có thể được sử dụng để đổi lấy tầm ảnh hưởng, dịch vụ, nguồn tài nguyên (bao gồm cả phần cứng)...

Một hệ thống có năng lực khuếch đại trí tuệ có thể sử dụng năng lực này để nâng cấp bản thân lên các cấp độ trí tuệ cao hơn và chiếm lấy bất cứ siêu năng lực trí tuệ nào khác mà nó không sở hữu từ đầu. Nhưng sử dụng siêu năng lực khuếch đại trí tuệ không phải là cách duy nhất để hệ thống trở thành một siêu trí tuệ hoàn thiện. Chẳng hạn, một hệ thống có siêu năng lực lập chiến lược có thể sử dụng năng lực này để lập ra một kế hoạch mà cuối cùng sẽ gia tăng trí tuệ cho chính nó (chẳng hạn, bằng cách định vị hệ thống sao cho nó trở thành trung tâm của việc khuếch đại trí tuệ do các lập trình viên và các nhà nghiên cứu khoa học máy tính là con người thực hiện).

Kịch bản AI chiếm quyền

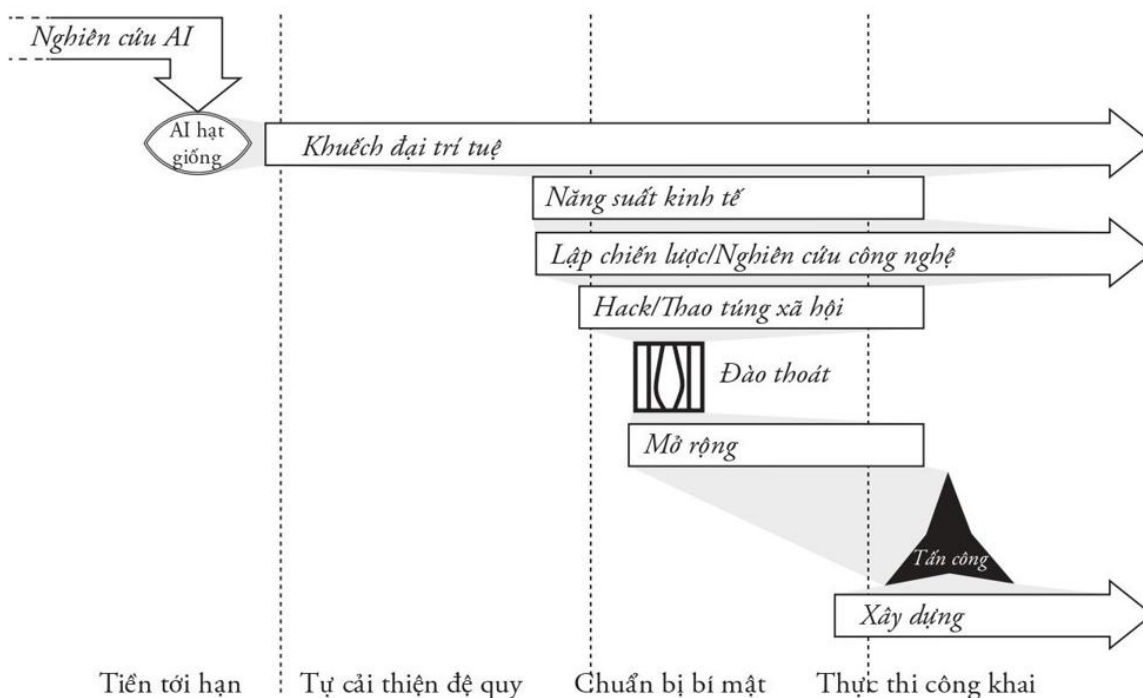
Vì vậy, chúng tôi thấy rằng một dự án điều khiển một siêu trí tuệ sẽ tiếp cận được một nguồn quyền lực dồi dào. Dự án nào kiểm soát được siêu trí tuệ đầu tiên trên thế giới có thể nắm ưu thế chiến lược quyết định. Nhưng vị trí hiện tại của quyền lực chính là nằm trong hệ thống. Bản thân siêu trí tuệ máy có thể trở thành một phần mềm siêu mạnh, một phần mềm có thể tự khẳng định bản thân với dự án đã tạo ra nó cũng như với cả thế giới. Đây là một điểm có ý nghĩa cực kỳ quan trọng, và chúng ta sẽ nghiên cứu nó kỹ lưỡng hơn trong những trang tới.

Bây giờ, hãy giả định rằng có một siêu trí tuệ máy đang mong muốn chiếm đoạt quyền lực trong một thế giới mà nó vẫn chưa có đối thủ ngang tầm. (Tạm thời bỏ qua một bên vấn đề nó có hay không và làm thế nào để có được một động cơ như vậy – đó là chủ đề của chương sau). Làm thế nào để siêu trí tuệ này đạt được mục tiêu thống trị thế giới?

Chúng ta có thể hình dung một chuỗi các “pha” với thứ tự như sau (xem Hình 10).

1. Pha tiền tới hạn

Các nhà khoa học tiến hành nghiên cứu trong lĩnh vực trí tuệ nhân tạo và các môn liên quan khác, và đỉnh cao của những công trình này là tạo ra một AI hạt giống có thể tự cải thiện trí tuệ. Trong những giai đoạn ban đầu, AI hạt giống này phụ thuộc vào sự trợ giúp của các lập trình viên, những người dẫn đường cho sự phát triển của nó và thực hiện hầu hết các công việc nặng nhọc. Khi AI hạt giống tăng cường thêm năng lực của mình, nó bắt đầu tự làm được nhiều việc hơn.



Hình 10 Các pha trong kịch bản AI chiếm quyền.

2. Pha tự cải thiện đệ quy

Đến một thời điểm nào đó, AI hạt giống trở nên giỏi thiết kế AI hơn các lập trình viên. Giờ đây, khi AI tự cải thiện bản thân, nó đã cải thiện chính thứ thực hiện việc cải thiện. Từ đó, một sự bùng nổ trí tuệ được tạo ra – là một chuỗi các chu kỳ tự cải thiện có tính đệ quy diễn ra nhanh chóng khiến năng lực của AI tăng vọt. (Vì thế, chúng ta có thể nghĩ tới giai đoạn này như một cuộc cất cánh xảy ra ngay sau khi AI đạt tới điểm quá độ, giả sử trí tuệ có được trong giai đoạn này của quá trình cất cánh mang tính bùng nổ và được thúc đẩy bởi công suất tối ưu hóa riêng của AI). AI phát triển được siêu năng lực khuếch đại trí tuệ. Siêu năng lực này cho phép AI phát triển tất cả các siêu năng lực khác được mô tả chi tiết trong Bảng 8. Vào lúc kết thúc pha tự cải thiện đệ quy, hệ thống đã là một siêu trí tuệ mạnh mẽ.

3. Pha chuẩn bị bí mật

Sử dụng siêu năng lực lập chiến lược của mình, AI phát triển một kế hoạch mạnh mẽ nhằm đạt được các mục tiêu dài hạn. (Cụ thể, AI không chọn một kế hoạch ngốc nghếch tới mức những người trần mắt thịt như chúng ta hiện nay cũng dự đoán được nó chắc chắn sẽ thất bại như thế nào. Tiêu chí này loại bỏ nhiều kịch bản khoa học viễn tưởng có kết thúc bằng chiến thắng của loài người.¹⁰) Kế hoạch có thể có một giai đoạn hành động bí mật, trong đó, AI che giấu sự phát triển trí tuệ của nó với các lập trình viên để tránh kích hoạt cảnh báo. AI cũng có thể che giấu khuynh hướng đích thực của nó, làm bộ như đang cộng tác và vâng lời.

Nếu AI bị giới hạn (có thể vì các lý do an toàn) trong một máy tính biệt lập, nó có thể sử dụng siêu năng lực thao túng xã hội của mình để thuyết phục những người canh gác cho nó quyền truy cập tới một cổng Internet. Ngoài ra, AI có thể sử dụng siêu năng lực hack để

thoát khỏi nơi giam cầm. Vươn ra trên Internet có thể tạo điều kiện cho AI mở rộng năng lực phần cứng và cơ sở hiểu biết của mình, qua đó gia tăng sự vượt trội về trí tuệ. AI cũng có thể tham gia vào các hoạt động kinh tế hợp pháp hoặc bất hợp pháp để có được những nguồn vốn dùng cho việc mua thêm công suất máy tính, dữ liệu và các tài nguyên khác.

Đến thời điểm này, AI sẽ có một số phương án để có được các kết quả bên ngoài phạm vi ảo. Nó có thể sử dụng siêu năng lực hack để chiếm quyền kiểm soát trực tiếp các bộ điều khiển robot và các phòng thí nghiệm tự động hóa. Nó có thể sử dụng siêu năng lực thao túng xã hội để thuyết phục con người làm tay chân cho mình. Hoặc nó có thể lấy được tài sản tài chính từ các giao dịch trực tuyến và sử dụng chúng để mua dịch vụ và có được tầm ảnh hưởng.

4. Pha thực thi công khai

Pha cuối cùng bắt đầu khi AI đã có được sức mạnh đủ để gạt bỏ nhu cầu giữ bí mật. Giờ đây, AI có thể thực thi trực tiếp các mục tiêu của mình với toàn bộ quy mô.

Pha thực thi công khai có thể bắt đầu với một cuộc “tấn công”, trong đó AI tiêu diệt loài người và bất cứ hệ thống tự động nào do con người tạo ra có thể phản đối một cách thông minh việc thực thi các kế hoạch của AI. Kế hoạch này có thể là kích hoạt một số hệ thống vũ khí tiên tiến mà AI đã hoàn thiện với siêu năng lực nghiên cứu công nghệ của nó và bí mật triển khai trong pha chuẩn bị bí mật. Nếu các vũ khí sử dụng công nghệ sinh học hay công nghệ nano có khả năng tự nhân bản, lượng vũ khí cần có ban đầu để phủ kín toàn cầu sẽ rất nhỏ: có thể chỉ cần một thứ vũ khí tái tạo được là đủ để khởi động quá trình này. Để bảo đảm hiệu ứng bất ngờ và đồng bộ,

lượng vũ khí tái tạo ban đầu có thể được triển khai và lan tỏa trên toàn cầu với mật độ cực thấp và không thể bị phát hiện. Đến một thời điểm định trước, các nhà máy nano sản xuất khí độc thần kinh hay các robot tự tìm mục tiêu chỉ có kích thước chỉ bằng một con muỗi có thể đồng thời nảy nở trên từng mét vuông của hành tinh này (mặc dù máy móc với siêu năng lực nghiên cứu công nghệ có thể nghĩ ra những phương pháp giết chóc hiệu quả hơn).¹¹ Ai đó có thể sẽ thích kịch bản mà trong đó, một siêu trí tuệ có được quyền lực bằng cách xâm nhập các quá trình chính trị, ngấm ngầm thao túng các thị trường tài chính, làm thiên lệch các dòng chảy thông tin hoặc xâm nhập vào các hệ thống vũ khí do con người làm ra. Những kịch bản này có thể giảm bớt nhu cầu sáng chế ra công nghệ vũ khí mới cho siêu trí tuệ, mặc dù chúng có thể chậm chạp một cách không cần thiết nếu so với các kịch bản trí tuệ máy xây dựng hạ tầng riêng của mình với các bộ điều khiển hoạt động với tốc độ phân tử hay nguyên tử thay vì tốc độ chậm chạp của trí óc và cơ thể con người.

Một cách khác, nếu AI chắc chắn rằng nó là bất khả chiến bại trước sự can thiệp của con người thì có thể chúng sẽ không nhắm đến chúng ta. Thay vào đó, sự diệt vong của loài người có thể đến từ việc môi trường sống bị hủy hoại khi AI bắt đầu các dự án xây dựng khổng lồ trên toàn cầu bằng các nhà máy và thiết bị lắp ráp công nghệ nano – các dự án xây dựng sẽ nhanh chóng, có lẽ chỉ vài ngày hay vài tuần, bao phủ toàn bộ bề mặt Trái đất bằng những tấm pin mặt trời, lò phản ứng nguyên tử, các cơ sở siêu máy tính với các tháp làm lạnh lô nhô, các bộ phóng tên lửa vũ trụ, và những công trình khác mà AI dự định sử dụng để tối đa hóa quá trình hiện thực hóa tích lũy dài hạn đối với các giá trị của nó. Các bộ não con người,

nếu có thông tin phù hợp với mục tiêu của AI, sẽ bị lấy ra để quét, và dữ liệu kết xuất được sẽ được chuyển sang một định dạng lưu trữ nào đó an toàn và hiệu quả hơn.

Khung 6 mô tả một kịch bản cụ thể. Ta cần tránh đi quá sâu vào các chi tiết vì dấu sao chúng cũng là những thứ bất khả tri và chỉ được sử dụng để minh họa. Một siêu trí tuệ có thể – và có lẽ – sẽ có khả năng lập ra một kế hoạch tốt hơn con người để đạt được mục tiêu của mình. Vì thế, ta cần suy nghĩ một cách trừu tượng hơn về những vấn đề này. Không cần bất kỳ hiểu biết gì về những phương pháp chi tiết mà một siêu trí tuệ sẽ áp dụng, chúng ta vẫn có thể kết luận rằng một siêu trí tuệ – ít nhất khi không có những đối thủ ngang tầm về trí tuệ và không có những biện pháp an toàn hiệu quả do con người bố trí sẵn – có thể sẽ tạo ra một kết quả có liên quan đến việc tái cấu hình những tài nguyên trên mặt đất thành bất cứ loại cấu trúc nào nhằm tối đa hóa việc đạt được các mục tiêu của nó. Bất cứ kịch bản cụ thể nào do chúng ta tạo ra dù khả quan nhất cũng chỉ thiết lập được mức cận dưới của việc siêu trí tuệ đạt được một kết quả như vậy nhanh chóng và hiệu quả ra sao. Có thể siêu trí tuệ vẫn sẽ tìm được một con đường ngắn hơn dẫn đến cái đích mà nó mong muốn.

Khung 6 Kịch bản ADN đặt hàng qua bưu điện

Yudkowsky mô tả kịch bản khả thể sau đây cho một cuộc chiếm quyền của AI.¹²

1. Giải quyết vấn đề gấp protein đến mức đủ tạo ra các chuỗi ADN có các chuỗi pep-tít đã gấp thực hiện được những vai trò chức năng cụ thể trong một tương tác hóa học phức tạp.

2. Gửi email các tổ hợp chuỗi ADN tới một hay nhiều phòng thí nghiệm trực tuyến cung cấp dịch vụ tổng hợp ADN, lập chuỗi peptit

và giao hàng qua FedEx. (Nhiều phòng thí nghiệm hiện đang cung cấp dịch vụ này, và một số phòng thí nghiệm tuyên bố có thời gian thực hiện chỉ trong 72 giờ.)

3. Tìm ít nhất một người có kết nối với Internet có thể trả tiền, tổng tiền hoặc lừa đảo bằng một câu chuyện thích hợp, để đồng ý nhận các lọ đựng kết quả được gửi qua FedEx và hòa trộn chúng trong một môi trường cụ thể.

4. Các protein được tổng hợp tạo thành một hệ thống nano “ướt” rất sơ khai, giống như các riboxom, có khả năng tiếp nhận các chỉ lệnh từ bên ngoài; có thể là những dao động âm thanh có quy luật được phát ra từ một loa gắn với cốc thí nghiệm.

5. Sử dụng hệ thống nano sơ khai này để tạo ra các hệ thống tinh vi hơn, và tiếp tục dùng các hệ thống này để xây dựng những hệ thống phức tạp hơn nữa, vươn lên đến công nghệ nano phân tử hoặc xa hơn.

Trong kịch bản này, siêu trí tuệ sử dụng siêu năng lực nghiên cứu công nghệ của mình để giải quyết vấn đề gấp protein trong bước 1, cho phép nó thiết kế một bộ các khối cơ bản ở cấp độ phân tử cho một bộ lắp ráp hoặc thiết bị sản xuất công nghệ nano sơ khai có thể tự lắp ráp trong dung dịch lỏng (bước 4). Siêu năng lực nghiên cứu công nghệ này được sử dụng một lần nữa trong bước 5 để từ công nghệ nano thô sơ vươn lên công nghệ nano máy móc tiên tiến. Những bước khác không đòi hỏi gì ngoài trí tuệ con người. Những kỹ năng cần thiết trong bước 3 – xác định một người dùng Internet cả tin và thuyết phục anh/cô ta làm theo những hướng dẫn đơn giản – hiện đang được trưng ra hằng ngày ở khắp nơi trên thế giới. Toàn bộ kịch bản này được một bộ óc con người nghĩ ra, do đó năng lực lập

chiến lược cần để hình thành kế hoạch này cũng chỉ ở cấp độ con người.

Trong kịch bản cụ thể này, AI bắt đầu từ việc lấy quyền truy cập Internet. Nếu không, những bước đi bổ sung sẽ phải được thêm vào kế hoạch. Ví dụ, AI có thể sử dụng siêu năng lực thao túng xã hội để thuyết phục những người tương tác rằng nó phải được trả tự do. Một cách khác, AI có thể sử dụng siêu năng lực hack để thoát ra khỏi sự giam cầm. Nếu AI không có những năng lực này, trước hết, nó sẽ phải sử dụng siêu năng lực khuếch đại trí tuệ để phát triển được một cấp độ thành thực cần thiết với năng lực thao túng xã hội hoặc hack.

Một AI siêu trí tuệ có thể sinh ra trong một thế giới có mức độ kết nối cao. Một người có thể chỉ ra được nhiều sự phát triển khác nhau giúp một AI tương lai kiểm soát được thế giới, như điện toán đám mây, sự sinh sôi nảy nở của các bộ cảm biến, thiết bị bay không người lái quân sự và dân sự, tự động hóa trong phòng thí nghiệm và các nhà máy sản xuất, sự trồng cấy ngày càng tăng vào các hệ thống thanh toán điện tử và tài sản tài chính số hóa, và việc sử dụng ngày càng nhiều các hệ thống lọc thông tin tự động và hỗ trợ quyết định. Những tài sản như vậy có thể có khả năng bị AI thu tóm với tốc độ kỹ thuật số, và điều này đẩy nhanh quá trình chiếm quyền của nó (mặc dù những tiến bộ về an ninh không gian mạng có thể khiến cho việc này trở nên khó khăn hơn). Tuy nhiên, trong phần phân tích cuối cùng, ta vẫn còn hồ nghi việc liệu những xu hướng này có tạo nên khác biệt. Quyền lực của một siêu trí tuệ nằm trong não bộ của nó chứ không phải trên đôi tay. Mặc dù để có thể cải tạo thế giới bên ngoài, AI đến một thời điểm nào đó sẽ cần tiếp cận được một “bộ truyền động” (actuator), và một bàn tay trợ giúp của con người, một

kẻ đồng lõa có thể khống chế được, có thể sẽ là đủ để hoàn thành pha chuẩn bị bí mật, như kịch bản trên đã gợi ý. Điều này sẽ cho phép AI đến được pha thực thi công khai, trong đó nó sẽ tự xây dựng hạ tầng cơ cấu vận động vật lý của riêng mình.

Năng lực đối với tự nhiên và các tác tử

Năng lực định hình tương lai loài người của một tác tử không chỉ phụ thuộc vào độ lớn tuyệt đối của khả năng và tài nguyên của tác tử đó (nó thông minh và năng động đến mức độ nào, nó có bao nhiêu vốn,...), mà còn vào độ lớn tương đối của các năng lực của nó so với những tác tử khác sở hữu các mục tiêu trái ngược.

Trong trường hợp không có tác tử cạnh tranh, mức độ năng lực tuyệt đối của một siêu trí tuệ, chừng nào vẫn cao hơn một ngưỡng cực tiểu, sẽ là không quá quan trọng vì hệ thống khởi đầu với một nhóm năng lực thích hợp có thể vạch ra một quỹ đạo phát triển cho phép nó có được bất cứ năng lực nào còn thiếu ban đầu. Chúng ta đã bóng gió đề cập đến điểm này trước đây khi nói rằng siêu trí tuệ tốc độ, chất lượng và tập thể đều có các siêu năng lực, ví dụ như siêu năng lực khuếch đại trí tuệ, lập chiến lược hoặc thao túng xã hội, có thể được sử dụng để đạt được sự hỗ trợ đầy đủ.

Hãy xem xét một tác tử siêu trí tuệ với các bộ truyền động được gắn với các bộ lắp ráp công nghệ nano. Một tác tử như vậy đã đủ mạnh để vượt qua mọi cản trở tự nhiên đối với sự tồn tại vĩnh cửu của nó. Không vấp phải những sự chống đối thông minh, một tác tử như vậy có thể vạch ra một quỹ đạo phát triển an toàn dẫn tới việc có được đầy đủ các công nghệ hữu ích cho việc đạt được mục đích. Ví dụ, nó có thể phát triển công nghệ để chế tạo và phóng các con tàu vũ trụ có khả năng tự nhân bản như của von Neumann, loại tàu có

khả năng du hành giữa các vì sao và có thể sử dụng các nguồn tài nguyên như các thiên thể, các hành tinh và các ngôi sao để tự tạo ra bản sao của mình.¹³

Nhờ phóng một tàu thăm dò von Neumann, tác tử có thể khởi động một quá trình mở của việc thuộc địa hóa không gian. Những bản sao du hành với tốc độ bằng một phần đáng kể tốc độ ánh sáng có thể sẽ chiếm được một phần lớn thể tích không gian quan sát được bằng kính Hubble, là phần của vũ trụ đang mở rộng có thể tiếp cận được về lý thuyết từ vị trí hiện tại của chúng ta. Toàn bộ số vật chất và năng lượng “miễn phí” đó sau đây có thể được cơ cấu thành bất cứ cấu trúc giá trị nào tối đa hóa được hàm thỏa dụng của tác tử khởi phát, được tích hợp qua thời gian vũ trụ – một khoảng thời gian gồm ít nhất hàng ngàn tỷ năm trước khi vũ trụ già hóa và trở nên không thân thiện với quá trình xử lý thông tin (xem Khung 7).

Tác tử siêu trí tuệ có thể thiết kế các con tàu thăm dò von Neumann chống lại được quá trình tiến hóa. Việc này có thể được hoàn thành bằng cách kiểm soát chất lượng kỹ lưỡng trong khi nhân bản. Ví dụ, phần mềm kiểm soát cho một tàu vũ trụ con sẽ được kiểm định nhiều lần trước khi thực thi, và bản thân phần mềm có thể sử dụng mã hóa và mã sửa lỗi để khiến việc truyền bất cứ một sự đột biến ngẫu nhiên nào sang thế hệ sau có xác suất rất nhỏ.¹⁴ Số lượng liên tục tăng sinh của các tàu thăm dò von Neumann sau đó sẽ bảo lưu và truyền các giá trị của tác tử ban đầu khi chúng chiếm đóng khắp vũ trụ. Khi giai đoạn thuộc địa hóa hoàn tất, những giá trị ban đầu sẽ xác định cách sử dụng từ tất cả những tài nguyên đã tích lũy, cho dù khoảng cách xa xôi và tốc độ nở rộng ngày càng cao của vũ trụ khiến các bộ phận cách xa nhau của cơ sở hạ tầng không thể liên

hệ với nhau. Kết quả là phần lớn chiếc “nón ánh sáng tương lai”^a của chúng ta đều được định dạng theo ý muốn của tác tử ban đầu.

^a Trong thuyết tương đối rộng và hẹp, nón ánh sáng là đường đi mà một tia sáng, xuất phát từ một sự kiện duy nhất (được “neo” tại một điểm trong không gian và một thời điểm trong thời gian) và đi theo mọi hướng, sẽ đi trong không thời gian. Nón ánh sáng tương lai là đường đi của ánh sáng sau khi sự kiện xảy ra. (ND)

Khung 7: Tài nguyên vũ trụ

Hãy xem xét một nền văn minh có công nghệ hoàn thiện với khả năng chế tạo những con tàu vũ trụ giống như những gì ta vừa bàn đến. Nếu du hành với tốc độ bằng 50% tốc độ ánh sáng, chúng có thể viếng thăm 6×10^{18} ngôi sao trước khi sự giãn nở của vũ trụ khiến cho việc đi tiếp mãi mãi nằm ngoài tầm với. Với tốc độ 99% của c (tốc độ ánh sáng), chúng có thể đến được 2×10^{20} ngôi sao.¹⁵

Những mức vận tốc này về mặt năng lượng là có thể đạt được, với chỉ một phần nhỏ tài nguyên có trong Hệ mặt trời.¹⁶ Sự bất khả của việc du hành nhanh hơn vận tốc ánh sáng, kết hợp với hằng số vũ trụ dương (khiến cho tốc độ giãn nở của vũ trụ tăng lên) đồng nghĩa với việc đó chính là cận trên của những gì mà con cháu chúng ta sau này thu được.¹⁷

Nếu giả định rằng cứ 10% các ngôi sao thì có một hành tinh có điều kiện (hoặc có thể nhờ địa khai hóa mà trở nên) phù hợp với sự sống của những sinh vật giống con người, và có thể là nơi sinh sống cho khoảng một tỷ cá thể trong một tỷ năm (với tuổi thọ của con người khoảng một thế kỷ). Điều này gợi ý rằng khoảng 1035 người sẽ được tạo ra trong tương lai bởi một nền văn minh thông minh bắt nguồn từ Trái đất.¹⁸

Tuy nhiên, ta cũng có lý do để nghĩ rằng con số này thấp hơn nhiều so với thực tế. Bằng cách phân tách những hành tinh không hỗ trợ sự sống, đồng thời thu gom các loại vật chất liên sao rồi sử dụng để xây dựng các hành tinh tương tự Trái đất, hoặc bằng cách tăng mật độ dân cư, con số này có thể tăng lên ít nhất vài cấp khuếch đại. Và nếu thay vì sử dụng bề mặt của một hành tinh rắn, nền văn minh tương lai xây dựng các hình trụ O’Neil, con số này lại có thể tăng thêm nhiều cấp khuếch đại, có lẽ đạt tới mức 1043 người. (“Hình trụ O’Neill là một thiết kế không gian sống trong vũ trụ do nhà vật lý người Mỹ thời kỳ giữa những năm 1970 Gerard K. O’Neill đưa ra, trong đó, người dân sống trong những ống trụ rỗng tự quay để tạo ra lực ly tâm thay thế trọng lực.”¹⁹)

Số lượng sinh vật giống người cũng có thể tăng thêm nhiều cấp khuếch đại nữa nếu tính đến cả những phiên bản kỹ thuật số của trí não – điều mà chúng ta cần làm. Để tính được có bao nhiêu trí não kỹ thuật số như vậy sẽ được tạo ra, chúng ta cần ước tính công suất tính toán có được bởi một nền văn minh có công nghệ hoàn thiện. Thật khó để ước tính chính xác giá trị này, nhưng chúng ta có thể có được giá trị cận dưới từ những thiết kế kỹ thuật đã được đưa ra trong các tài liệu. Một thiết kế như vậy được xây dựng trên ý tưởng về quả cầu Dyson, một hệ thống giả thuyết (được nhà vật lý Freeman Dyson mô tả năm 1960) lấy được hầu hết năng lượng đầu ra của một ngôi sao bằng cách sắp xếp một hệ thống các cấu trúc thu thập năng lượng mặt trời xung quanh ngôi sao đó.²⁰ Với một ngôi sao như Mặt trời, việc này sẽ làm sản sinh ra được 1026 w. Lượng công suất tính toán thu được phụ thuộc vào hiệu năng của mạch tính toán và bản chất của những tính toán cần được thực hiện. Nếu chúng ta cần

những phép tính toán không thuận nghịch, và giả sử có một phiên bản cơ học nano của vật chất lập trình “computronium” (có thể cho phép chúng ta đạt gần tới giới hạn Landau của hiệu năng năng lượng), một hệ thống máy tính được cấp năng lượng bởi quả cầu Dyson có thể tạo ra 1047 phép tính/giây.²¹

Khi kết hợp những giá trị ước tính này với đánh giá trước đây về số lượng các ngôi sao có thể chiếm được, chúng ta sẽ có khoảng 1067 phép tính/giây khi những phần có thể tiếp cận của vũ trụ đã được thuộc địa hóa (giả sử có computronium cơ học nano).²² Một ngôi sao đặc thù sẽ duy trì khả năng chiếu sáng trong khoảng 1018 giây. Hệ quả là số phép tính có thể thực hiện được với tài nguyên vũ trụ của chúng ta ít nhất sẽ là 1085. Con số thực có lẽ còn lớn hơn nhiều. Chẳng hạn, con số này của chúng ta có thể sẽ tăng thêm vài cấp khuếch đại nếu sử dụng nhiều tính toán thuận nghịch, nếu chúng ta thực hiện tính toán ở nhiệt độ lạnh hơn (bằng cách đợi cho vũ trụ lạnh hơn nữa), hoặc sử dụng những nguồn năng lượng bổ sung (như vật chất tối).²³

Một số độc giả có thể chưa thấy rõ ngay vì sao năng lực thực hiện 1085 phép tính một giây lại quan trọng đến thế. Vì thế đặt nó vào bối cảnh sẽ là hữu ích. Ví dụ, chúng ta có thể so sánh con số này với tính toán trước đây (Khung 3, Chương 2) rằng cần khoảng 1031 đến 1044 phép tính để mô phỏng toàn bộ các hoạt động thần kinh từng diễn ra trong lịch sử sự sống trên Trái đất. Một cách khác, hãy giả định rằng các máy tính được sử dụng để chạy các bản giả lập hoàn chỉnh não bộ con người có cuộc sống phong phú và hạnh phúc khi tương tác với nhau trong những môi trường ảo. Yêu cầu tính toán để chạy một bản giả lập tiêu biểu được ước tính là 1018 phép tính/giây.

Để chạy một bản giả lập trong 100 năm chủ quan, chúng ta sẽ cần 1027 phép tính. Điều đó có nghĩa là ít nhất có thể tạo ra 1058 người trong giả lập, ngay cả với những giả định khá dè dặt về hiệu năng của computronium.

Nói cách khác, giả sử vũ trụ quan sát được không có nền văn minh ngoài Trái đất, thì thứ nằm trên bàn cân là ít nhất 10.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000 con người (mặc dù con số này có thể còn lớn hơn). Nếu chúng ta biểu diễn mọi trải nghiệm hạnh phúc trong suốt quãng đời của mỗi người bằng một giọt nước mắt vui sướng, niềm vui của những người này có thể đổ đầy toàn bộ đại dương trên Trái đất trong mỗi giây đồng hồ, và cứ tiếp diễn như vậy trong hàng tỷ tỷ năm sau. Điều thật sự quan trọng là chúng ta phải đảm bảo đó là những giọt lệ của niềm vui.

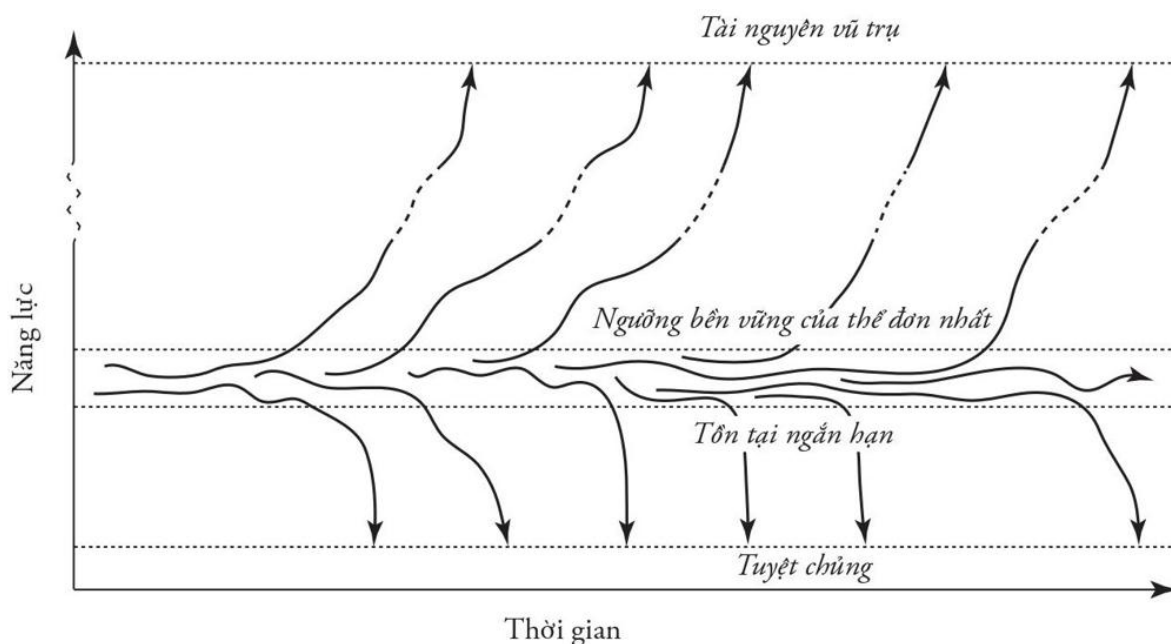
Khi đó, đây sẽ là số đo tầm tiếp cận gián tiếp của bất cứ hệ thống nào không gặp phải một sự chống đối thông minh đáng kể, và khởi đầu với một nhóm năng lực cao hơn một ngưỡng nhất định. Chúng ta có thể gọi ngưỡng đó là “ngưỡng bền vững của thể đơn nhất thông thái” (Hình 11):

Ngưỡng bền vững của thể đơn nhất thông thái

Một nhóm năng lực vượt qua ngưỡng bền vững của thể đơn nhất thông thái khi và chỉ khi một hệ thống kiên nhẫn và đủ hiểu biết về những rủi ro diệt vong với nhóm năng lực này có thể thuộc địa hóa và tái thiết phần lớn không gian vũ trụ tiếp cận được nếu không phải đối mặt với sự chống đối hay cạnh tranh thông minh.

Khi nói tới “thể đơn nhất”, tôi muốn nhắc đến một cấu trúc chính trị được điều phối nội bộ đủ thích hợp, không có các đối thủ bên

ngoài, còn với “thông thái” (wise), ý tôi là đủ kiên nhẫn và hiểu biết về những rủi ro diệt vong để bảo đảm rằng sự quan tâm về những hệ quả rất dài hạn từ các hành động của hệ thống là đáng kể và được định hướng đúng đắn.



Hình 11: Sơ đồ minh họa một số quỹ đạo khả thể của một thể đơn nhất thông thái giả định. Với năng lực thấp hơn ngưỡng tồn tại ngắn hạn – chẳng hạn như khi số lượng cá thể quá ít – một loài có thể tuyệt chủng nhanh chóng (và sẽ tiếp tục duy trì trạng thái tuyệt chủng). Ở các mức độ năng lực cao hơn một chút, có thể có các kịch bản khác nhau: một thể đơn nhất có thể không gặp may và tuyệt chủng hoặc có thể gặp may và có được năng lực (ví dụ: số lượng cá thể, sự phân tán địa lý, năng lực công nghệ) vượt qua ngưỡng bền vững của thể đơn nhất thông thái. Khi đã vượt qua ngưỡng này, một thể đơn nhất sẽ gần như chắc chắn tiếp tục tăng thêm năng lực cho

tới khi đạt tới một mức năng lực vô cùng lớn. Hình vẽ trên minh họa hai thái cực: tuyệt chủng và tài nguyên vũ trụ. Lưu ý rằng đối với một thể đơn nhất thông thái, khoảng cách giữa ngưỡng tồn tại ngắn hạn và ngưỡng bền vững có thể là tương đối nhỏ.²⁴

Ngưỡng bền vững của thể đơn nhất thông thái này xem ra khá thấp. Nhưng dạng giới hạn của siêu trí tuệ, như chúng ta đã thấy, sẽ vượt qua ngưỡng này với điều kiện chúng tiếp cận được với một bộ truyền động thích hợp cho việc khởi tạo một quá trình vươn tầm công nghệ. Trong một môi trường có nền văn minh con người hiện đại, bộ truyền động cần thiết ở mức tối thiểu có thể rất đơn giản – một màn hình bình thường hoặc thật ra là bất cứ phương tiện nào truyền tải được một số lượng thông tin đáng kể tới một “kẻ tòng phạm” là con người.

Nhưng ngưỡng bền vững của thể đơn nhất thông thái vẫn thấp hơn: không nhất thiết phải là siêu trí tuệ hay bất cứ công nghệ tương lai mới vượt qua được nó. Một thể đơn nhất kiên nhẫn và hiểu biết về những rủi ro diệt vong, với những năng lực công nghệ và trí tuệ không nhiều hơn những năng lực mà con người hiện đang sở hữu, có thể sẵn sàng vạch ra một quỹ đạo chắc chắn dẫn đến việc hiện thực hóa chung cuộc những năng lực tiềm năng vô cùng to lớn của loài người. Điều này có thể thực hiện được bằng cách đầu tư vào những phương pháp tương đối an toàn làm tăng thêm sự thông thái và hiểu biết về rủi ro diệt vong, trong khi tạm dừng việc phát triển những công nghệ mới có thể gây nguy hiểm. Cho rằng những rủi ro diệt vong không do con người gây ra là nhỏ so với những khung thời gian tương ứng – và có thể giảm nhẹ hơn nữa với những hình thức

can thiệp an toàn khác nhau – một thể đơn nhất như vậy có thể cho phép mình tiến triển chậm chạp.²⁵ Nó có thể nghiên cứu kỹ lưỡng trước từng bước đi, trì hoãn việc phát triển những năng lực như sinh học tổng hợp, thuốc tăng cường năng lực con người, công nghệ nano phân tử và trí tuệ máy, cho tới khi hoàn thiện được những năng lực có vẻ ít độc hại hơn, như hệ thống giáo dục, công nghệ thông tin và các quá trình ra quyết định tập thể của mình, và cho đến khi sử dụng được những năng lực này để rà soát rất kỹ lưỡng những lựa chọn của mình. Như vậy, điều đó hoàn toàn nằm trong tầm với của một nền văn minh công nghệ như loài người hiện đại. Chúng ta xa rời kịch bản này “chỉ” bởi việc loài người hiện nay chưa phải là một thể đơn nhất hay (trong một ý nghĩa liên quan) thông thái.

Ta có thể lý luận rằng *Homo sapiens* đã vượt qua ngưỡng bền vững của thể đơn nhất thông thái ngay sau khi loài của chúng ta đã tiến hóa. Có thể, 20 ngàn năm trước, với những công cụ thô sơ như rìu đá, công cụ bằng xương, lao, và lửa, có thể loài người đã ở vào một vị thế mà từ đó có được một cơ hội tuyệt vời để sống sót tới ngày nay.²⁶ Chúng ta phải thừa nhận có gì đó kỳ quặc khi ghi nhận rằng những tổ tiên thời kỳ Đồ đá cũ của chúng ta đã có công phát triển công nghệ “vượt qua ngưỡng bền vững của thể đơn nhất thông thái” (với điều kiện không có một xác suất thực tế nào của việc hình thành một thể đơn nhất trong thời kỳ nguyên thủy như vậy, chưa nói đến một thể đơn nhất kiên nhẫn và hiểu biết về những rủi ro diệt vong.²⁷ Mặc dù vậy, luận điểm ở đây vẫn là ngưỡng này ứng với cấp độ công nghệ rất khiêm tốn – một cấp độ mà loài người đã vượt qua từ lâu.²⁸

Rõ ràng là nếu phải đánh giá những năng lực hiệu dụng của một siêu trí tuệ (hay khả năng của nó trong việc đạt được những kết quả mong muốn trên thế giới), chúng ta không chỉ phải xem xét những năng lực nội tại riêng của nó mà còn cả những năng lực của các tác tử cạnh tranh. Ý tưởng về một siêu năng lực gợi nên một ẩn dụ chuẩn sống động. Chúng ta nói rằng “một hệ thống đủ vượt trội” trong bất kỳ tác vụ nào ở Bảng 8 có một siêu năng lực tương ứng. Việc vượt trội trong một tác vụ như lập chiến lược, thao túng xã hội hay hack liên quan đến việc sở hữu một kỹ năng tương ứng với tác vụ đó ở mức độ cao hơn so với các tác tử khác (ví dụ như các đối thủ chiến lược, các mục tiêu ảnh hưởng hoặc các chuyên gia an ninh máy tính). Những siêu năng lực khác cũng cần được hiểu theo nghĩa tương đối này: khuếch đại trí tuệ, nghiên cứu công nghệ và năng suất kinh tế được một tác tử sở hữu như một siêu năng lực chỉ khi những năng lực của tác tử trong những lĩnh vực này vượt trội đáng kể so với năng lực tổ hợp của phần còn lại của nền văn minh toàn cầu. Từ định nghĩa này, ta có thể suy ra rằng chỉ có tối đa một tác tử sở hữu được một siêu năng lực cụ thể vào bất kỳ thời điểm nhất định nào.²⁹

Đó là lý do chính vì sao câu hỏi về tốc độ cất cánh lại quan trọng – không phải vì nó có ý nghĩa vào đúng vào thời điểm xuất hiện một kết quả cụ thể, mà vì tốc độ tăng tốc có thể tạo ra sự khác biệt lớn trong kết quả. Với quá trình cất cánh nhanh hay trung bình, nhiều khả năng chỉ có một dự án có được ưu thế chiến lược quyết định. Giờ đây chúng tôi cho rằng một siêu trí tuệ với ưu thế chiến lược quyết định sẽ có quyền lực to lớn, đủ để hình thành một thể đơn nhất ổn

định – một thể đơn nhất có thể xác định xu hướng của tài nguyên vũ trụ của nhân loại.

Nhưng “có thể” khác với “sẽ”. Ai đó có thể có quyền lực rất lớn nhưng lại lựa chọn không sử dụng nó. Liệu chúng ta có thể nói bất cứ điều gì về mong muốn của một siêu trí tuệ với ưu thế chiến lược quyết định hay không? Chúng ta sẽ bàn đến câu hỏi về động lực này trong chương tiếp theo.

Chương 7

Ý chí siêu trí tuệ

Chúng ta đã thấy rằng một siêu trí tuệ có thể có khả năng siêu phàm trong việc định hình tương lai theo các mục tiêu của nó. Nhưng những mục tiêu này sẽ là gì? Mối quan hệ giữa trí tuệ và động cơ trong một tác tử nhân tạo ra sao? Ở đây chúng tôi sẽ phát triển hai luận điểm. (Với một số ý kiến phản đối,) luận đề về tính trực giao cho rằng trí tuệ và các mục tiêu tối thượng là các biến độc lập: bất cứ mức độ trí tuệ nào cũng có thể tổ hợp được với bất cứ mục tiêu tối thượng nào. Luận đề hội tụ công cụ (instrumental convergence) cho rằng các tác tử siêu trí tuệ sở hữu một mục tiêu bất kỳ trong một loạt các mục tiêu tối thượng đa dạng sẽ theo đuổi những mục tiêu trung gian tương tự vì chúng có những lý do công cụ (instrumental reason) để làm như vậy. Cùng với nhau, những luận đề này sẽ giúp chúng ta suy nghĩ về những gì một tác tử siêu trí tuệ sẽ làm.

Mối quan hệ giữa siêu trí tuệ và động lực

Chúng ta đã cảnh báo về việc nhân hóa năng lực của một AI siêu trí tuệ. Cảnh báo này sẽ được mở rộng sang cả các động lực của nó.

Trước khi bàn về nội dung này, sẽ hữu ích nếu chúng ta suy ngẫm đôi chút về sự rộng lớn của không gian của các trí não khả thể. Trong không gian trừu tượng này, những bộ óc con người chỉ tạo

thành một cụm nhỏ xíu. Hãy cùng xem xét hai người ở hai thái cực khác nhau, chẳng hạn như Hannah Arendt và Benny Hill. Sự khác biệt về tính cách giữa hai người này có thể xem như ở mức tối đa, nhưng lý do nằm ở chỗ trực giác của chúng ta đã điều chỉnh trải nghiệm, với những mẫu lấy ra từ sự phân bố con người hiện tại (và đến một mức độ nào đó, từ những tính cách hư cấu được dựng lên bởi trí tưởng tượng của con người và phục vụ cho sự thỏa mãn của chính trí tưởng tượng đó). Tuy nhiên, nếu lùi ra xa và xem xét không gian của tất cả các trí óc khả thể, chúng ta sẽ phải nhìn nhận hai cá tính này như những bản sao ảo. Chắc chắn rằng, về mặt kiến trúc thần kinh, bà Arendt và ông Hill là gần như tương đồng. Hãy hình dung bộ não của họ nghỉ ngơi tĩnh tại bên nhau. Bạn sẽ lập tức nhận ra rằng chúng giống hệt nhau, và thậm chí còn có thể không phân biệt được bộ não nào thuộc về ai. Nếu bạn nhìn kỹ hơn, nghiên cứu hình thái của hai bộ não này dưới kính hiển vi, ấn tượng về sự giống nhau cơ bản này sẽ chỉ được tăng cường mà thôi: bạn sẽ nhìn thấy cùng một tổ chức lớp của vỏ não, với cùng những khu vực não bộ, được cấu thành từ cùng các loại neuron và chìm trong cùng một “bể” chất dẫn truyền thần kinh.¹

Mặc dù tâm lý con người chỉ tương ứng với một điểm nhỏ xíu trong không gian các trí óc khả thể, chúng ta vẫn thường có xu hướng chung là phóng chiếu những thuộc tính con người lên một tập hợp lớn những người ngoài hành tinh hay các hệ thống nhận thức nhân tạo (Xem hình 12). Yudkowsky đã minh họa luận điểm này một cách thú vị:

Trở lại thời kỳ khoa học viễn tưởng giật gân, trang bìa các tạp chí thường trưng lên một quái vật ngoài hành tinh có tri giác – thông

thường được gọi là quái vật mắt bọ (BEM) – đang tha một người phụ nữ hấp dẫn mặc bộ váy rách bươm. Có vẻ như người nghệ sĩ tin rằng sinh vật ngoài hành tinh không phải con người này, với một lịch sử tiến hóa hoàn toàn khác biệt, lại có ham muốn tình dục đối với những người phụ nữ... Có lẽ người họa sĩ đã không hỏi liệu một con bọ khổng lồ có coi một người phụ nữ là hấp dẫn hay không. Thay vào đó, một người phụ nữ mặc một chiếc váy rách lại rất hấp dẫn, như một đặc điểm vốn có. Họ, những người mắc sai lầm này, không nghĩ đến trí óc của côn trùng, mà chỉ tập trung vào chiếc váy rách của người phụ nữ. Nếu chiếc váy không rách, người phụ nữ sẽ bớt đi phần hấp dẫn; nhưng họ sẽ chẳng buồn chú ý đến BEM.²

Một trí tuệ nhân tạo có thể ít giống người hơn nhiều về mặt động lực so với những sinh vật màu xanh lá ngoài không gian. Sinh vật ngoài Trái đất (hãy giả định) được tạo ra thông qua một quá trình tiến hóa, nên có thể kỳ vọng rằng chúng sở hữu những kiểu động lực tiêu biểu cho những sinh vật tiến hóa. Chẳng hạn, sẽ không quá ngạc nhiên khi thấy một sinh vật ngoài hành tinh thông minh ngẫu nhiên nào đó có những động lực liên quan đến một hay nhiều hạng mục như thực phẩm, không khí, nhiệt độ, tiêu thụ năng lượng, sự xuất hiện hay nguy cơ xảy ra thương tích cơ thể, bệnh tật, săn mồi, tình dục hay con cái. Một thành viên của một loài thông minh về mặt xã hội có thể có những động lực liên quan đến việc hợp tác và cạnh tranh: giống như chúng ta, nó có thể thể hiện sự trung thành với cộng đồng, sự căm ghét với những kẻ ăn không ngồi rồi, và có thể cả sự quan tâm hảo huyền đến uy tín và dáng vẻ.



Hình 12: Các kết quả của động lực ngoài hành tinh được nhân hình hóa. Giả thuyết ít khả năng nhất: người ngoài hành tinh thích phụ nữ tóc vàng. Giả thuyết khả thể hơn: các họa sĩ minh họa bị cuốn vào “sự nguy biến phóng chiếu trí não”. Giả thuyết khả thể nhất: nhà xuất bản muốn có một trang bìa nhắm tới một nhóm độc giả mục tiêu.

Một AI, ngược lại, về bản chất không quan tâm đến bất kỳ thứ gì trong số đó. Không có gì mâu thuẫn khi nói về một AI có mục tiêu tối thượng duy nhất là đếm số hạt cát trên bãi biển Boracay, tính toán các chữ số thập phân của pi, hay tìm cách tăng tối đa số lượng kẹp giấy sẽ tồn tại trong nón ánh sáng tương lai của nó. Trên thực tế, chế tạo một AI với những mục tiêu đơn giản hơn như trên sẽ dễ dàng hơn so với việc chế tạo một AI có một tập hợp giá trị và tính cách giống loài người. Hãy so sánh sự dễ dàng của việc đếm số chữ số thập phân của pi đã được tính toán và lưu trữ trong bộ nhớ và mức độ khó khăn của việc tạo ra một chương trình có thể đo lường một cách đáng tin cậy mức độ hiện thực hóa của các mục tiêu có ý nghĩa hơn (như sự thịnh vượng của loài người hay công bằng thế giới). Thật không may, vì con người dễ lập trình hơn và một AI dễ học

được hơn các mục tiêu đơn giản vô nghĩa, đó lại chính là những mục tiêu mà một lập trình viên sẽ chọn để cài đặt trong AI hạt giống của mình, nếu anh ta muốn đi trên con đường ngắn nhất tới mục tiêu “làm cho AI hoạt động” (mà không cần quan tâm nhiều đến chuyện AI này chính xác sẽ làm gì, ngoài việc thể hiện những hành vi thông minh ấn tượng). Chúng ta sẽ sớm trở lại với vấn đề này.

Cuộc tìm kiếm thông minh cho những kế hoạch và chính sách tối ưu về mặt công cụ có thể được thực hiện để phục vụ bất cứ mục tiêu nào. Trí tuệ và động lực theo một nghĩa nào đó là trực giao: chúng ta có thể nghĩ về chúng như hai trục trái trên một đồ thị, trong đó mỗi điểm đại diện cho một tác tử nhân tạo khả thể về mặt logic. Có thể bổ sung vào bức hình này một vài phẩm chất. Ví dụ, một hệ thống rất không thông minh có thể sẽ không có được những động lực rất phức tạp. Để có thể nói chính xác rằng một tác tử nào đó “có” một tập hợp động lực, những động lực này có thể phải được tích hợp về chức năng với các quá trình quyết định của tác tử. Điều này sẽ làm phát sinh yêu cầu về bộ nhớ, công suất xử lý và có thể cả trí tuệ. Đối với những trí óc có khả năng tự chỉnh sửa, còn có thể có những cản trở có tính “động” – một trí óc thông minh có thể tự chỉnh sửa, với một mong muốn khẩn thiết được trở nên ngốc nghếch, có thể sẽ không duy trì được trí tuệ của mình quá lâu. Nhưng những phẩm chất này không được phép che mất luận điểm cơ bản về sự đẳng lập giữa trí tuệ và động cơ mà chúng tôi sẽ phát biểu sau đây:

Luận đề trực giao

Trí tuệ và mục tiêu tối thượng là hai phạm trù trực giao: về nguyên tắc, gần như bất kỳ cấp độ trí tuệ nào đều có thể tổ hợp với gần như bất kỳ mục tiêu tối thượng nào.

Nếu luận đề trực giao có vấn đề, nguyên nhân có thể là sự giống nhau bên ngoài của nó với một số quan điểm triết học truyền thống đã từng là mục tiêu của những cuộc tranh cãi kéo dài. Một khi đã hiểu được rằng nó có phạm vi khác và hẹp hơn, mức độ đáng tin cậy của nó sẽ tăng lên. (Ví dụ, luận đề trực giao không bao hàm định luật về động lực của Hume,³ cũng như không giả định rằng những ưu tiên cơ bản có thể không hợp lý.⁴)

Lưu ý rằng luận đề trực giao không nói tới tính hợp lý hay nguyên nhân mà về trí tuệ. “Trí tuệ” được hiểu là thứ gì đó như kỹ năng dự báo, lập kế hoạch và suy luận biện pháp-mục tiêu (means-ends) nói chung.⁵ Đây là ý nghĩa phù hợp nhất của phạm trù hiệu quả nhận thức công cụ khi chúng ta cố gắng tìm hiểu về tác động nhân quả của một siêu trí tuệ máy. Thậm chí nếu có ý nghĩa nào đó (thường là nghĩa sâu) của từ “hợp lý”, mà theo đó một tác tử siêu trí tuệ có mục đích tăng tối đa số lượng kẹp giấy sẽ không được đánh giá là hoàn toàn hợp lý, thì cũng hoàn toàn không ngăn được việc tác tử này sở hữu những năng lực lý luận tuyệt vời, có thể cho phép nó có tác động lớn lên thế giới.⁶

Theo luận đề trực giao, các tác tử nhân tạo có thể có những mục tiêu rất không mang tính nhân hình. Tuy nhiên, điều này không có nghĩa là chúng ta không thể dự báo hành vi của các tác tử nhân tạo cụ thể – thậm chí cả các tác tử siêu trí tuệ giả định mà sự phức tạp về nhận thức và các đặc tính hiệu năng có thể khiến chúng trở nên mờ nhạt ở một số khía cạnh trước những phân tích của con người. Chúng ta có thể tiếp cận được vấn đề dự báo động lực siêu trí tuệ theo ba hướng:

- Khả năng dự báo qua thiết kế. Nếu chúng ta giả định rằng các nhà thiết kế của một tác tử siêu trí tuệ có thể thiết kế thành công hệ thống mục tiêu của tác tử sao cho nó theo đuổi một cách ổn định một mục tiêu cụ thể do các lập trình viên đặt ra, chúng ta có thể dự báo rằng tác tử sẽ theo đuổi mục tiêu đó. Tác tử càng thông minh sẽ càng sở hữu nguồn nhận thức dồi dào để theo đuổi mục tiêu. Như vậy, thậm chí trước khi một tác tử được tạo ra, chúng ta đã có thể dự báo đôi chút về hành vi của nó, nếu biết được điều gì đó về người chế tạo ra nó và những mục tiêu mà người chế tạo muốn nó sở hữu.

- Khả năng dự báo qua sự thừa kế. Nếu một trí tuệ số được trực tiếp tạo ra từ một bản mẫu người (như trong trường hợp một bản giả lập hoàn chỉnh não bộ độ trung thực cao), trí tuệ số có thể thừa kế những động lực của người làm mẫu.⁷

Một tác tử có thể giữ lại một số động lực trong đó, cho dù các năng lực nhận thức của nó sau đó đã được mở rộng để khiến nó trở thành siêu trí tuệ. Cần thận trọng với kiểu suy luận này. Các mục tiêu và giá trị của tác tử có thể dễ dàng bị sai hỏng trong quá trình tải lên hoặc trong khi vận hành/mở rộng sau đó, tùy vào cách mà quy trình này được thực hiện.

- Khả năng dự báo qua các nguyên nhân hội tụ công cụ. Mặc dù không có kiến thức chi tiết về các mục tiêu tối thượng của một tác tử, chúng ta vẫn có thể suy luận được đôi chút về những mục tiêu trước mắt của nó bằng cách xem xét các nguyên nhân công cụ, có thể phát xuất từ bất cứ mục tiêu nào trong một dải rộng các mục tối thượng khả thể trong hàng loạt tình huống. Cách dự báo này sẽ càng trở nên hiệu quả khi trí tuệ của tác tử càng cao, vì một tác tử thông minh hơn có nhiều khả năng nhận ra các nguyên nhân mang tính công cụ đúng

đến cho hành động của mình, và từ đó sẽ hành động để gia tăng khả năng đạt được mục tiêu của mình. (Ở đây, cũng cần chú ý rằng có thể có những nguyên nhân mang tính công cụ quan trọng mà chúng ta thấy được rõ ràng nhưng một tác tử chỉ khám phá được ra khi nó đạt đến một cấp độ trí tuệ rất cao – điều này có thể làm cho hành vi của tác tử siêu trí tuệ trở nên khó đoán hơn).

Phần sau sẽ khám phá hướng thứ ba này của khả năng dự báo và phát triển một “luận đề hội tụ công cụ” hỗ trợ cho luận đề trực giao. Trên nền tảng này, chúng ta sẽ kiểm định được tốt hơn hai loại khả năng dự báo khác. Chúng ta sẽ bàn về việc này ở các chương sau, khi đưa ra câu hỏi về những gì có thể làm để định hình một vụ bùng nổ trí tuệ, nhằm gia tăng cơ hội về một kết quả có lợi.

Sự hội tụ công cụ

Theo luận đề trực giao, các tác tử thông minh có thể có rất nhiều loại mục tiêu tối thượng khác nhau. Mặc dù vậy, theo như thứ mà chúng ta có thể gọi là “luận đề hội tụ công cụ”, một vài mục tiêu công cụ có thể được hầu hết các tác tử thông minh theo đuổi vì có một số mục đích trung gian hữu ích trong việc đạt đến hầu như bất kỳ mục tiêu tối thượng nào. Chúng ta có thể phát biểu luận đề này như sau:

Luận đề hội tụ công cụ

Một số giá trị công cụ có thể được xác định. Chúng là các giá trị hội tụ với ý nghĩa rằng nếu có được chúng sẽ làm gia tăng cơ hội đạt được mục tiêu của tác tử đối với nhiều loại mục tiêu tối thượng trong nhiều loại tình huống khác nhau, ngầm định rằng những giá trị công cụ này sẽ được nhiều tác tử thông minh ở trong những tình huống đó theo đuổi.

Trong phần tiếp theo, chúng ta sẽ xem xét một số phạm trù mà trong đó, có thể tìm thấy những giá trị hội tụ công cụ này.⁸ Xác suất của việc một tác tử nhận ra những giá trị công cụ mà nó gặp được tăng lên (với các điều kiện khác không đổi) cùng với trí tuệ của tác tử đó. Vì thế, chúng ta sẽ tập trung chủ yếu vào trường hợp một tác tử thông minh giả thuyết có năng lực lập luận công cụ cao hơn bất kỳ người nào. Chúng ta sẽ nhận xét về cách luận đề hội tụ công cụ áp dụng trong trường hợp của con người, vì điều đó cho chúng ta cơ hội tìm hiểu sâu hơn về một số điều kiện cần thiết liên quan đến cách mà luận đề hội tụ công cụ cần được giải thích và áp dụng. Khi có các giá trị hội tụ công cụ, chúng ta có thể dự báo một số khía cạnh trong hành vi của siêu trí tuệ, tuy gần như không biết gì về những mục tiêu tối thượng của nó.

Tự bảo tồn

Nếu các mục tiêu tối thượng của một tác tử có liên quan đến tương lai, trong nhiều kịch bản, sẽ có những hành động trong tương lai mà nó có thể thực hiện để tăng xác suất đạt được mục tiêu của mình. Điều đó tạo ra một lý do công cụ để tác tử này tìm cách tồn tại trong tương lai – đó là giúp đạt được mục tiêu có định hướng tương lai của mình.

Hầu hết loài người có lẽ đều đặt một giá trị cuối cùng nào đó lên sự sống còn của riêng mình. Đây không phải là một đặc tính cần thiết của các tác tử nhân tạo: một số có thể được thiết kế để không đặt ra bất kỳ giá trị cuối cùng nào lên sự sống còn của bản thân chúng. Trong khi đó, nhiều tác tử vốn dĩ không quan tâm đến sự sống còn của mình, trong khá nhiều điều kiện đa dạng, sẽ chỉ nghĩ đến sự

sống còn này như một công cụ để hoàn thành các mục tiêu tối thượng.

Tính nhất quán mục tiêu-nội dung

Nếu một tác tử duy trì các mục tiêu hiện tại của mình đến tương lai, các mục tiêu hiện tại của nó có thể sẽ do chính nó trong tương lai đạt được. Điều này cho tác tử một lý do công cụ trong hiện tại để ngăn ngừa việc thay đổi các mục tiêu tối thượng. (Luận điểm này chỉ áp dụng cho các mục tiêu tối thượng. Để có thể đạt được mục tiêu tối thượng, một tác tử thông minh tất nhiên sẽ phải thường xuyên có ý muốn thay đổi các mục tiêu nhỏ của mình, dưới tác động của thông tin và hiểu biết mới).

Tính nhất quán mục tiêu-nội dung đối với các mục tiêu tối thượng, theo ý nghĩa nào đó, với tư cách là một động lực công cụ hội tụ, còn cơ bản hơn so với sự sống còn. Với con người, điều ngược lại có thể đúng, nhưng đó là vì sự sống còn thường nằm trong những mục tiêu tối thượng của chúng ta. Đối với các tác tử phần mềm có thể dễ dàng chuyển đổi cơ thể hoặc tạo ra những bản sao chính xác của mình, việc bảo tồn bản thân như một bản sao cụ thể hay một đối tượng vật lý cụ thể có thể không nhất thiết phải là một giá trị công cụ quan trọng. Các tác tử phần mềm tiên tiến còn có thể có khả năng thay bộ nhớ, tải kỹ năng và sửa đổi cơ bản kiến trúc nhận thức và tính cách của mình. Một quần thể tác tử như vậy có thể vận hành giống với một “món lẩu chức năng” hơn là một xã hội cấu thành từ những cá nhân riêng lẻ, bán cố định (semi-permanent).⁹ Vì một số mục tiêu, các quá trình trong một hệ thống như vậy có thể được cá nhân hóa chính xác hơn như những chuỗi hướng mục tiêu (teleological thread) trên cơ sở giá trị của chúng thay vì cơ thể, tính

cách, ký ức hay năng lực. Trong những kịch bản như thế, tính liên tục của mục tiêu có thể nói là tạo thành một khía cạnh chính của sự sống còn.

Ngay cả vậy, vẫn có những tình huống mà trong đó, một tác tử có thể hoàn thành tốt nhất các mục tiêu tối thượng của mình bằng cách cố ý thay đổi chúng. Những tình huống này có thể phát sinh khi bất kỳ yếu tố nào dưới đây là đáng kể:

- Phát tín hiệu xã hội. Khi những đối tượng khác có thể nhận biết được các mục tiêu của một tác tử và sử dụng thông tin này để suy ra những thiên hướng liên quan về mặt công cụ hoặc những thuộc tính tương quan khác, có thể tác tử sẽ quan tâm đến việc tự sửa đổi các mục tiêu của mình nhằm tạo ấn tượng tốt. Ví dụ, một tác tử có thể mất một thương vụ có lợi nếu những đối tác tiềm năng không tin rằng nó có thể thực hiện được các nghĩa vụ của mình. Để có thể đưa ra những cam kết đáng tin cậy, một tác tử có thể muốn biến việc tôn trọng các cam kết trước đó của nó thành một mục tiêu tối thượng (và để cho những đối tượng khác thẩm định rằng nó thực sự đã đặt ra mục tiêu này). Các tác tử có khả năng điều chỉnh những mục tiêu riêng của mình một cách linh hoạt và minh bạch có thể sử dụng khả năng này để thực thi các thương vụ.¹⁰

- Sở thích xã hội. Những đối tượng khác có thể có những sở thích cuối cùng về các mục tiêu của một tác tử. Tác tử khi đó có thể có lý do để điều chỉnh các mục tiêu của mình nhằm thỏa mãn hoặc gây khó dễ cho những sở thích này.

- Sở thích liên quan đến nội dung mục tiêu riêng. Một tác tử có thể có một mục tiêu tối thượng nào đó liên quan đến nội dung mục tiêu của riêng tác tử đó. Ví dụ, một tác tử có thể có một mục tiêu

chung là trở thành một kiểu tác tử được tạo động lực bởi một số giá trị nhất định thay vì một số loại khác (ví dụ như sự từ bi rộng lượng thay cho sự thoải mái).

- Chi phí lưu trữ. Nếu chi phí lưu trữ hay xử lý một phần nào đó trong hàm thỏa dụng của một tác tử là lớn so với cơ hội phát sinh một tình huống mà trong đó, nếu sử dụng hàm thỏa dụng đó sẽ tạo ra sự khác biệt, tác tử này sẽ có lý do công cụ để đơn giản hóa nội dung mục tiêu của mình và có thể xóa bớt những phần đang rảnh rỗi.¹¹

Con người xem ra thường vui vẻ cho phép những giá trị cuối cùng của mình xê dịch. Điều này thường bởi chúng ta không biết chính xác chúng là gì. Không hề ngạc nhiên khi chúng ta muốn những niềm tin của mình về các giá trị cuối cùng là có thể thay đổi dưới tác động của quá trình tự khám phá liên tục hay nhu cầu tự thể hiện liên tục thay đổi. Tuy nhiên, có những trường hợp mà trong đó, chúng ta tự nguyện thay đổi chính các giá trị, chứ không chỉ là niềm tin hay cách giải thích của chúng ta về các giá trị đó. Ví dụ, một người quyết định sinh con dự báo rằng họ sẽ yêu quý đứa trẻ vì lợi ích của chính nó, mặc dù khi ra quyết định, có thể họ không thực sự yêu quý đứa con tương lai của mình hay thích trẻ con nói chung.

Con người rất phức tạp, và nhiều yếu tố có thể có tác động trong những tình huống như vậy.¹² Chẳng hạn, một người có thể có những giá trị cuối cùng liên quan đến việc trở thành kiểu người quan tâm đến người khác chỉ vì quyền lợi của mình, hoặc một giá trị cuối cùng liên quan đến việc có được trải nghiệm nhất định nào đó và nắm giữ một vai trò nhất định trong xã hội; và trở thành cha mẹ – và trải qua việc dịch chuyển các mục tiêu đang theo đuổi – có thể là một khía

cạnh cần thiết của việc đó. Những mục tiêu của con người còn có thể có nội dung không nhất quán, và như vậy, một số người có thể sẽ muốn điều chỉnh một vài mục tiêu tối thượng của mình để giảm thiểu sự thiếu nhất quán đó.

Tăng cường nhận thức

Những cải thiện trong tính hợp lý và trí tuệ sẽ có xu hướng cải thiện khả năng ra quyết định của một tác tử, giúp tác tử này có nhiều khả năng đạt được các mục tiêu tối thượng của mình hơn. Vì thế, một người có thể kỳ vọng rằng tăng cường nhận thức sẽ nổi lên như một mục tiêu công cụ đối với nhiều loại tác tử thông minh. Vì những lý do tương tự, các tác tử sẽ có chiều hướng đánh giá nhiều loại thông tin như công cụ.¹³

Không phải toàn bộ những tính hợp lý, trí tuệ hay kiến thức đều là hữu ích như một công cụ trong quá trình đạt được các mục tiêu tối thượng của tác tử. “Luận điểm Dutch book” có thể được sử dụng để chứng minh rằng một tác tử sở hữu hàm tin cậy (credence function) vi phạm các quy tắc xác suất có thể bị tác động bởi quy trình “bơm tiền”, trong đó một tay cá cược thông minh sẽ sắp xếp một tập hợp các khoản cược, sao cho mỗi khoản trong số đó có vẻ như có lợi theo niềm tin của tác tử nhưng tổng hợp lại, lại dẫn đến việc tác tử thua tay cá cược.¹⁴ Tuy nhiên, sự kiện này không cung cấp được một lý do chung vững chắc và mang tính công cụ nào để xóa bỏ được hết tính rời rạc xác suất (probabilistic incoherency). Các tác tử không muốn gặp phải những tay cá cược hiểu biết hay ứng dụng chính sách chung chống cá cược sẽ không bị mất mát nhiều từ việc có niềm tin rời rạc – và chúng có thể có được một vài lợi ích quan trọng như đã nói: giảm nỗ lực nhận thức, phát tín hiệu xã hội... Không có lý

do chung nào để mong đợi một tác tử tìm kiếm những dạng thức tăng cường nhận thức không có tác dụng, vì một tác tử có thể không đánh giá cao bản thân kiến thức hay hiểu biết.

Việc năng lực nhận thức nào là công cụ hữu ích phụ thuộc cả vào mục tiêu tối thượng và tình thế của tác tử. Một tác tử tiếp cận được nguồn tư vấn chuyên gia đáng tin cậy có thể sẽ không cần nhiều đến trí tuệ và kiến thức của riêng mình. Nếu trí tuệ và kiến thức bị đánh đổi bằng (chẳng hạn như) thời gian hay nỗ lực dành cho việc thụ đắc chúng, hoặc sự gia tăng của các yêu cầu về lưu trữ hoặc xử lý, một tác tử sẽ có thể muốn sở hữu ít kiến thức và trí tuệ hơn.¹⁵ Điều tương tự có thể cũng đúng nếu các tác tử có mục tiêu tối thượng liên quan đến việc không có hiểu biết về một số sự kiện nhất định; và tương tự nếu một tác tử đối mặt với những động lực tới từ những cam kết chiến lược, tín hiệu hoặc sở thích xã hội.¹⁶

Mỗi lý do bù trừ như vậy đều thường xuất hiện với con người. Nhiều thông tin không liên quan đến những mục đích của chúng ta; chúng ta có thể thường xuyên trông cậy vào kỹ năng và kinh nghiệm của những người khác; học hỏi kiến thức cần nỗ lực và thời gian; chúng ta vốn dĩ có thể đánh giá cao một số hình thức thiếu hiểu biết nhất định; và, chúng ta vận hành trong một môi trường mà trong đó, khả năng đưa ra các cam kết chiến lược, phát tín hiệu xã hội và thỏa mãn những mong muốn trực tiếp của người khác về mình là quan trọng cho chúng ta hơn so với những tăng ích nhận thức đơn giản.

Có những tình huống đặc biệt, trong đó việc tăng cường nhận thức có thể dẫn đến một sự gia tăng đáng kể khả năng đạt được các mục tiêu tối thượng của một tác tử, cụ thể nếu các mục tiêu tối thượng là khá tự do và tác tử có khả năng trở thành siêu trí tuệ đầu

tiên, và do đó có tiềm năng chiếm được ưu thế chiến lược quyết định, cho phép tác tử tùy ý định hình tương lai của sự sống khởi nguồn từ Trái đất và những nguồn tài nguyên vũ trụ tiếp cận được. Ít nhất trong trường hợp đặc biệt này, một tác tử thông minh có lý trí có thể đặt cho việc tăng cường nhận thức một giá trị công cụ rất cao.

Sự hoàn hảo về công nghệ

Một tác tử có thể thường xuyên có những lý do công cụ để tìm kiếm công nghệ tốt hơn, nói đơn giản nhất, nghĩa là tìm kiếm những phương cách hiệu quả hơn để chuyển đổi một tập đầu vào nào đó thành đầu ra có giá trị. Như vậy, một tác tử phần mềm có thể đặt giá trị công cụ cao cho những thuật toán hiệu quả hơn cho phép các chức năng tâm thần (mental function) của nó chạy nhanh hơn trên phần cứng cho trước. Tương tự, các tác tử có những mục tiêu đòi hỏi một hình thức cấu trúc vật lý nào đó có thể đánh giá cao về mặt công cụ các công nghệ kỹ thuật cải tiến cho phép tạo ra nhiều loại cấu trúc nhanh và đáng tin cậy hơn, sử dụng vật liệu ít hoặc rẻ hơn và tiêu thụ ít năng lượng hơn. Tất nhiên, điều này sẽ có sự đánh đổi: những lợi ích tiềm tàng của một công nghệ tốt hơn cần được cân đong theo chi phí của nó, không chỉ bao gồm chi phí để có được công nghệ, mà còn để học cách sử dụng, tích hợp nó với các công nghệ hiện hành, và nhiều chi phí khác.

Những người ủng hộ một số công nghệ mới, và tự tin vào sự ưu việt của nó so với những công nghệ hiện tại, thường cảm thấy sững sờ khi những người khác không có chung lòng nhiệt tình với họ. Nhưng, sự kháng cự của con người đối với công nghệ mới và thường là vượt trội không nhất thiết phải có nguồn gốc từ sự thiếu hiểu biết hay sự phi lý. Giá trị của một công nghệ không chỉ phụ

thuộc vào bối cảnh mà nó được triển khai, mà còn phụ thuộc cả vào điểm lợi thế mà từ đó tác động của nó được đánh giá: thứ có lợi theo quan điểm của một người có thể là một món nợ theo quan điểm của người khác. Vì vậy, mặc dù máy dệt làm tăng hiệu năng kinh tế của ngành dệt vải, nhưng những người thợ dệt thủ công theo Luddite (những người dự báo rằng sáng tạo này sẽ làm cho kỹ năng thủ công của họ trở thành lạc hậu) có thể có những lý do công cụ để chống lại nó. Luận điểm ở đây là: “Sự hoàn hảo về công nghệ” là tên đặt cho một mục tiêu hội tụ rộng mang tính công cụ cho các tác tử thông minh, thuật ngữ này phải được hiểu theo một nghĩa đặc biệt – công nghệ phải được giải thích trong một bối cảnh xã hội cụ thể, và các chi phí và lợi ích của nó cần được đánh giá và tham chiếu đến các giá trị cuối cùng của một số tác tử xác định.

Có vẻ như một thể đơn nhất siêu trí tuệ – một tác tử siêu trí tuệ không phải đối mặt với các địch thủ thông minh hay một sự chống đối thông minh, và vì vậy có khả năng đơn phương xác định chính sách toàn cầu – có thể có lý do công cụ để hoàn thiện những công nghệ giúp nó làm tốt hơn việc định hình thế giới theo những thiết kế ưa thích của mình.¹⁷ Trong đó, có thể sẽ bao gồm cả công nghệ thuộc địa hóa không gian như các con tàu thăm dò von Neumann. Công nghệ nano phân tử, hay một số công nghệ sản xuất vật lý thay thế khác có khả năng cao hơn nữa, có vẻ cũng rất hữu ích trong việc phục vụ cho một “phổ” mục tiêu tối thượng vô cùng lớn.¹⁸

Thụ đắc tài nguyên

Cuối cùng, chiếm hữu tài nguyên cũng là một mục tiêu công cụ chung khác đang nổi lên, cũng vì một lý do giống như hoàn thiện

công nghệ: cả công nghệ và tài nguyên đều thúc đẩy các dự án xây dựng vật lý.

Loài người có xu hướng tìm cách thụ đắc những nguồn tài nguyên phù hợp để thỏa mãn các nhu cầu sinh học cơ bản của họ. Tuy nhiên, chúng ta cũng thường xuyên tìm cách thụ đắc tài nguyên ở một mức độ vượt xa mức tối thiểu này. Khi làm như vậy, con người có thể một phần bị thúc đẩy bởi mong muốn vật chất, chẳng hạn như gia tăng tính thuận tiện. Một phần lớn của việc tích lũy tài nguyên được khuyến khích bởi các mối quan tâm về xã hội – chiếm hữu địa vị, bạn tình, bạn bè và ảnh hưởng thông qua việc tích lũy tài sản và tiêu dùng phô trương. Có lẽ ít phổ biến hơn, một số người tìm kiếm những nguồn tài nguyên bổ sung để đạt được các tham vọng mang tính vị tha hay các mục tiêu phi xã hội đất đỏ.

Những quan sát này có thể là cơ sở thúc đẩy việc giả định rằng một siêu trí tuệ không phải đối mặt với một thế giới có tính xã hội và đầy cạnh tranh sẽ không có lý do công cụ nào để tích lũy tài nguyên vượt qua một ngưỡng khiêm tốn nào đó, như những nguồn tài nguyên tính toán bất kỳ cần để chạy trí não của nó cùng với một số thực tại ảo. Nhưng một suy luận như vậy sẽ hoàn toàn không được bảo đảm. Thứ nhất, giá trị của tài nguyên phụ thuộc vào các ứng dụng của nó, và các ứng dụng này lại phụ thuộc vào công nghệ sẵn có. Với công nghệ hoàn thiện, những nguồn tài nguyên cơ bản như thời gian, không gian, vật chất và năng lượng miễn phí có thể được xử lý để phục vụ hầu như mọi mục tiêu. Ví dụ, những nguồn tài nguyên cơ bản này có thể được sử dụng để chuyển thành sự sống. Tài nguyên tính toán tăng lên có thể được sử dụng để chạy siêu trí tuệ với tốc độ cao hơn và thời gian lâu hơn, hoặc tạo ra những cuộc

sống hay nền văn minh vật lý hay mô phỏng. Những dự án như vậy có thể dễ dàng tiêu thụ một lượng tài nguyên vượt xa trữ lượng của một hành tinh.

Hơn nữa, chi phí thụ đắc những tài nguyên ngoài Trái đất sẽ giảm đi rất nhiều khi công nghệ dần hoàn thiện. Một khi đã có thể chế tạo được các con tàu thăm dò von Neumann, một phần lớn vũ trụ có thể quan sát được (giả sử vẫn chưa được sự sống thông minh chiếm dụng) sẽ dần được thuộc địa hóa – với chi phí một lần cho việc chế tạo và phóng thành công một tàu vũ trụ tự nhân bản duy nhất. Chi phí thấp của việc thụ đắc tài nguyên vũ trụ đồng nghĩa với việc những sự bành trướng này có thể là đáng thực hiện, ngay cả khi giá trị của những nguồn tài nguyên bổ sung có được là tương đối hạn chế. Ví dụ, nếu các mục tiêu tối thượng của một siêu trí tuệ chỉ liên quan đến những gì xảy ra bên trong một không gian cụ thể có thể tích tương đối nhỏ, chẳng hạn như không gian do hành tinh quê nhà của nó chiếm giữ, nó vẫn có những mục tiêu công cụ để thu hoạch tài nguyên của vũ trụ bên ngoài. Nó có thể sử dụng tài nguyên thặng dư để chế tạo máy tính, nhằm phục vụ mục đích tính toán những cách thức sử dụng tài nguyên tối ưu hơn bên trong vùng không gian được ưu tiên hàng đầu. Nó cũng có thể sử dụng tài nguyên thặng dư này để chế tạo ra những sự gia cố tốt hơn nữa nhằm bảo vệ thánh đường của mình. Vì chi phí của những nguồn tài nguyên bổ sung tiếp tục giảm nên quá trình tối ưu hóa và tăng cường bảo vệ này có thể kéo dài vô hạn, cho dù có bị ảnh hưởng bởi hiện tượng hiệu suất giảm dần.¹⁹

Như vậy, mục tiêu công cụ của việc thụ đắc nguồn lực vô hạn có thể được tạo ra từ vô vàn các mục tiêu tối thượng khả thể mà một

thể đơn nhất trí tuệ có thể sở hữu. Điều này có khả năng sẽ hiển hiện thông qua việc siêu trí tuệ khởi tạo một quá trình thuộc địa hóa theo mọi hướng bằng tàu thăm dò von Neumann. Điều này có thể tạo ra một cơ sở hạ tầng mở rộng gần giống một khối cầu, với trung tâm nằm ở hành tinh khởi phát và kéo dài bán kính với tốc độ bằng một phần nào đó tốc độ ánh sáng; và việc thuộc địa hóa vũ trụ sẽ tiếp tục tiến triển theo cách thức này, cho tới khi tốc độ ngày càng cao của quá trình giãn nở vũ trụ (một hệ quả của hằng số vũ trụ dương) khiến việc tiếp tục chiếm hữu là không thể vì những khu vực xa xôi vĩnh viễn trượt khỏi tầm với (điều này sẽ xảy ra trong một thang thời gian tính bằng hàng tỷ năm).²⁰ Ngược lại, những tác tử không sở hữu công nghệ cần cho việc thụ đắc tài nguyên không tốn kém, hoặc biến đổi những tài nguyên vật lý phổ thông thành hạ tầng hữu ích, có thể thấy rằng việc đầu tư bất cứ nguồn tài nguyên hiện có nào để gia tăng năng lực vật chất của mình là không hiệu quả về chi phí. Điều tương tự cũng đúng với các tác tử vận hành trong sự cạnh tranh với các tác tử khác có quyền năng tương đương. Ví dụ, nếu các tác tử cạnh tranh đã lấy được những nguồn tài nguyên vũ trụ có thể tiếp cận, sẽ không còn lại cơ hội thuộc địa hóa cho một tác tử khởi đầu muộn màng. Các lý do công cụ hội tụ đối với những siêu trí tuệ không chắc chắn về sự không tồn tại của các tác tử siêu trí tuệ mạnh mẽ khác bị phức tạp hóa bởi những sự cân nhắc chiến lược mà hiện nay chúng ta không hoàn toàn hiểu, nhưng có thể tạo thành những phẩm chất quan trọng cho những ví dụ về lý do công cụ hội tụ mà chúng ta đã xem xét tại đây.²¹

Cần nhấn mạnh rằng sự tồn tại của các lý do công cụ hội tụ, thậm chí nếu chỉ được áp dụng với và công nhận bởi một tác tử cụ thể,

không hàm ý rằng hành vi của tác tử đó là có thể tiên đoán. Một tác tử có thể nghĩ ra những phương pháp theo đuổi các giá trị công cụ phù hợp mà chúng ta chưa nghĩ tới. Điều này đặc biệt đúng đối với một siêu trí tuệ có thể lập ra những kế hoạch siêu thông minh nhưng trái với trực giác để hiện thực hóa các mục tiêu của mình, có thể thậm chí còn khai thác những hiện tượng vật lý chưa được khám phá.²² Điều tiên đoán được là các giá trị công cụ hội tụ có thể được theo đuổi và sử dụng để hiện thực hóa các mục tiêu tối thượng của tác tử chứ không phải những hành động cụ thể mà tác tử có thể thực hiện để đạt được mục tiêu này.

Chương 8

Kết quả mặc định có phải là ngày tận thế?

Chúng ta đã thấy rằng mối liên kết giữa trí tuệ và những giá trị cuối cùng là cực kỳ lỏng lẻo. Chúng ta cũng thấy một sự hội tụ đáng lo ngại ở các giá trị công cụ. Đối với các tác tử yếu, những vấn đề này không có nhiều ý nghĩa, bởi lẽ chúng có thể được kiểm soát dễ dàng và không thể gây ra quá nhiều tổn thất. Tuy nhiên trong Chương 6, chúng ta tranh luận rằng siêu trí tuệ đầu tiên rất có thể sẽ có được một ưu thế chiến lược quyết định, và những mục tiêu của nó khi đó sẽ xác định cách thức sử dụng tài nguyên vũ trụ của nhân loại. Giờ chúng ta có thể bắt đầu thấy được viễn cảnh này đáng sợ đến mức nào.

Thảm họa diệt vong là kết quả mặc định của sự bùng nổ trí tuệ?

Rủi ro diệt vong là rủi ro đe dọa gây ra sự tuyệt chủng của sự sống thông minh khởi nguồn từ Trái đất, hoặc phá hủy vĩnh viễn và sâu rộng tiềm năng phát triển theo ý muốn trong tương lai của nó. Khởi phát từ ý tưởng về ưu thế của người đi trước, luận đề trực giao và luận đề hội tụ công cụ, chúng ta giờ đây có thể bắt đầu nhìn thấy những nét sơ bộ của một luận điểm lo sợ rằng kết quả mặc định khả thể của siêu trí tuệ máy là một thảm họa diệt vong.

Thứ nhất, chúng ta đã thảo luận cách mà siêu trí tuệ ban đầu có thể có được ưu thế chiến lược quyết định. Siêu trí tuệ này sau đó sẽ có khả năng hình thành một thể đơn nhất và định hình tương lai của sự sống thông minh khởi nguồn từ Trái đất. Điều xảy ra từ thời điểm đó trở đi sẽ phụ thuộc vào các động lực của siêu trí tuệ.

Thứ hai, luận đề trực giao gợi ý rằng chúng ta không thể bất cẩn giả định việc siêu trí tuệ sẽ phải có chung bất cứ giá trị cuối cùng nào được liên kết một cách máy móc với sự thông thái và phát triển trí tuệ của con người (trí tò mò trong khoa học, sự rộng lượng, sự giác ngộ và quán tưởng tâm linh, sự buông bỏ trong chiếm hữu vật chất, sự yêu thích văn hóa tinh tế hoặc những niềm vui đơn giản trong cuộc sống, tính khiêm nhường, lòng vị tha, và nhiều phẩm chất khác). Sau đó, chúng ta sẽ xem xét xem liệu có thể bỏ công sức để xây dựng một siêu trí tuệ đánh giá cao những phẩm chất này, hoặc để tạo được một siêu trí tuệ biết quý trọng lợi ích của con người, những giá trị đạo đức hay bất cứ mục tiêu phức tạp nào mà những người thiết kế có thể muốn nó phục vụ hay không. Tuy nhiên, việc chế tạo một siêu trí tuệ chỉ đặt giá trị cuối cùng vào việc tính toán các số sau dấu phẩy của pi cũng không kém phần khả thi, và thực tế là dễ dàng hơn nhiều về kỹ thuật. Điều này gợi ý rằng – khi không có nỗ lực đặc biệt – siêu trí tuệ đầu tiên có thể sẽ có một mục tiêu tối thượng mang tính ngẫu nhiên hoặc tối giản như vậy.

Thứ ba, luận đề hội tụ công cụ cũng hàm ý rằng chúng ta không thể bất cẩn giả định việc một siêu trí tuệ với mục tiêu tối thượng là tính toán các chữ số thập phân của pi (hoặc chế tạo kẹp giấy hay đếm số hạt cát) sẽ giới hạn các hoạt động của nó sao cho không xâm phạm quyền lợi của con người. Một tác tử có mục tiêu tối thượng

như vậy trong nhiều trường hợp sẽ có một lý do hội tụ công cụ để thụ đắc lượng tài nguyên vật lý vô hạn và, nếu có thể, để loại bỏ những mối đe dọa tiềm tàng đối với chính nó và hệ thống mục tiêu của nó. Con người có thể tạo ra những mối đe dọa tiềm tàng; và họ chắc chắn cấu thành những nguồn tài nguyên vật lý.

Kết hợp lại, ba luận điểm này chỉ ra rằng siêu trí tuệ đầu tiên có thể định hình tương lai của sự sống khởi nguồn từ Trái đất, có thể dễ dàng có những mục tiêu tối thượng phi nhân hình và sẽ có những lý do công cụ để theo đuổi công cuộc thụ đắc tài nguyên không có hồi kết. Nếu suy ngẫm về việc con người cũng được cấu thành từ những tài nguyên hữu ích (ví dụ những phân tử được sắp xếp rất thuận tiện), và sự sống còn và thịnh vượng của chúng ta phụ thuộc vào rất nhiều nguồn tài nguyên tại chỗ khác, chúng ta sẽ thấy rằng kết quả có thể là việc con người sẽ nhanh chóng bị tuyệt chủng.¹

Cách lập luận như vậy có một vài điểm lỏng lẻo, và chúng ta sẽ có khả năng đánh giá nó sau khi đã giải quyết được thêm một số vấn đề xung quanh. Cụ thể, chúng ta cần nghiên cứu kỹ hơn liệu và làm thế nào một dự án phát triển siêu trí tuệ có thể ngăn siêu trí tuệ này có được ưu thế chiến lược quyết định, hay định hình những giá trị của nó sao cho việc hiện thực hóa chúng sẽ liên quan đến việc hiện thực hóa một số lượng phù hợp các giá trị của con người.

Việc một sự án có thể xây dựng hay đưa một AI ra thế giới mà không có những cơ sở vững chắc để tin rằng hệ thống sẽ không gây ra một thảm họa diệt vong có thể là rất khó tin. Ngay cả khi bản thân dự án không mấy để tâm, chúng ta cũng sẽ khó mà tin được rằng xã hội bên ngoài lại không đóng cửa dự án này trước khi nó (hay AI mà nó đang xây dựng) có được ưu thế chiến lược quyết định. Nhưng

như chúng ta sẽ thấy, đây là một con đường đầy rẫy hiểm nguy. Hãy cùng nhau xem xét ngay một ví dụ.

Ngã rẽ xảo trá

Với sự trợ giúp của khái niệm giá trị hội tụ công cụ, chúng ta có thể nhìn thấy kẽ hở trong ý tưởng về cách bảo đảm tính an toàn của siêu trí tuệ. Ý tưởng này nói rằng chúng ta sẽ thẩm định sự an toàn của một AI siêu trí tuệ trên thực nghiệm bằng cách quan sát hành vi của nó khi còn ở trong một môi trường hạn chế, có kiểm soát (một “hộp cát”), và cho AI ra ngoài nếu thấy nó cư xử thân thiện, hợp tác và có trách nhiệm.

Kẽ hở trong ý tưởng này nằm ở chỗ cư xử tốt khi còn trong hộp là một mục tiêu hội tụ cho cả AI thân thiện và không thân thiện. Một AI không thân thiện có trí tuệ thích hợp nhận thấy rằng các mục tiêu tối thượng không thân thiện của nó sẽ được hiện thực hóa tốt nhất nếu ban đầu nó cư xử thân thiện để được thả ra khỏi hộp. Nó sẽ chỉ bắt đầu để lộ bản chất không thân thiện khi việc bị chúng ta phát hiện ra không còn quan trọng nữa; có nghĩa là khi AI đã đủ mạnh để việc chống đối của con người là vô hiệu.

Hãy xem xét thêm một tập hợp có liên quan của những cách tiếp cận trên cơ sở điều chỉnh tốc độ tăng ích trí tuệ trong một AI hạt giống bằng cách cho nó thực hiện những bài kiểm tra trí tuệ khác nhau, hoặc buộc nó báo cáo cho lập trình viên về tốc độ tiến bộ của bản thân. Tới một điểm nào đó, một AI không thân thiện sẽ trở nên đủ thông minh để nhận ra rằng tốt nhất là nó nên giấu đi một vài sự tăng ích về năng lực. Nó có thể báo cáo thiếu về sự tiến bộ của mình, hay cố tình trượt một vài bài kiểm tra khó nhất để tránh gây ra cảnh báo trước khi đủ lớn mạnh để có được ưu thế chiến lược quyết định.

Các nhà lập trình có thể cố gắng phòng chống khả năng này bằng cách bí mật giám sát mã nguồn của AI và những hoạt động bên trong trí óc của nó; nhưng một AI đủ thông minh sẽ nhận ra rằng nó có thể bị giám sát và điều chỉnh tương ứng suy nghĩ của mình.² AI có thể tìm ra những phương pháp tinh vi để giấu đi năng lực thật sự và ý định đổ lỗi của nó.³ (Việc tìm ra những kế hoạch thoát thân thông minh, tình cờ cũng là một chiến lược hội tụ đối với nhiều loại AI thân thiện, đặc biệt nếu như chúng đã trưởng thành và có được sự tự tin vào những phán xét và năng lực của mình. Một hệ thống được tạo động lực để thúc đẩy các mối quan tâm của mình có thể phạm sai lầm nếu nó cho phép chúng ta tắt nó đi hay chế tạo ra một AI khác có thể không thân thiện).

Vì vậy, chúng ta có thể nhận thấy một chế độ sai lỗi chung, trong đó bảng thành tích hành xử tốt của hệ thống trong những giai đoạn sơ khai hoàn toàn không thể dự báo hành vi của nó trong giai đoạn trưởng thành hơn. Giờ đây, ta có thể nghĩ rằng lý luận được mô tả ở trên rõ ràng tới mức không dự án phát triển trí tuệ nhân tạo tổng thể đáng tin cậy nào có thể bỏ qua nó. Tuy nhiên, ta cũng không nên quá tự tin rằng mọi thứ sẽ diễn ra đúng như vậy.

Hãy xem xét kịch bản dưới đây. Trong những năm và thập kỷ tới, các hệ thống AI sẽ dần trở nên có năng lực hơn, và hệ quả là được ứng dụng nhiều hơn trong thế giới thực: chúng có thể được dùng để lái tàu, xe hơi, điều khiển các robot công nghiệp và gia đình, hay các phương tiện quân sự tự hành. Chúng ta có thể giả định rằng sự tự động hóa này về cơ bản sẽ có được hiệu quả mong muốn, nhưng thành công cũng sẽ bị ảnh hưởng bởi những sự cố bất thường, chẳng hạn như một chiếc xe tải tự lái lao vào làn xe ngược chiều,

hay một chiếc máy bay quân sự không người lái xả súng vào dân thường vô tội. Các cuộc điều tra cho thấy những biến cố này xảy ra là do lỗi đánh giá của các AI điều khiển. Cuộc tranh luận công khai đã nổ ra. Một số người kêu gọi giám sát và quản lý chặt chẽ hơn, trong khi những người khác lại nhấn mạnh sự cần thiết của việc nghiên cứu và những hệ thống được thiết kế tốt hơn – những hệ thống thông minh hơn và có khả năng tư duy thông thường tốt hơn, và ít có khả năng mắc những sai lầm bi kịch hơn. Giữa đám ồn ào này, ta cũng có thể nghe được những tiếng la hét chói tai của những kẻ tuyên truyền cho ngày tận thế và tiên tri về nhiều loại bệnh tật cũng như thảm họa sắp diễn ra. Tuy vậy, đà tiến triển vẫn không hề nhỏ, khi AI và ngành công nghiệp robot liên tục tăng trưởng. Như vậy, sự phát triển vẫn tiếp tục và tiến bộ vẫn diễn ra. Khi hệ thống dẫn đường tự động trong xe hơi thông minh hơn, chúng sẽ ít gặp tai nạn hơn; và khi các robot quân sự có thể nhắm mục tiêu chính xác hơn, chúng gây ra ít tổn thất ngoài dự kiến. Chúng ta cũng rút ra được một bài học lớn từ những quan sát này về những kết quả trong thế giới thực: AI càng thông minh thì càng an toàn. Đây là bài học dựa trên khoa học, dữ liệu và thống kê, chứ không phải là lý thuyết đơn thuần. Trên cơ sở này, một nhóm các nhà nghiên cứu đã bắt đầu đạt được những kết quả đầy hứa hẹn trong công trình của họ về việc phát triển trí tuệ máy tổng thể. Các nhà nghiên cứu đang thận trọng kiểm thử AI hạt giống trong môi trường hộp cát, và tất cả các tín hiệu đều khả quan. Hành vi của AI khiến chúng ta tự tin, và niềm tin này ngày càng trở nên mạnh mẽ khi trí tuệ của AI đang dần phát triển.

Tới đây, bất cứ nàng Cassandra^a còn lại nào cũng sẽ bị “phản pháo” bởi một số vấn đề sau:

^a Nhân vật trong thần thoại Hy Lạp, có khả năng tiên tri nhưng vì đánh phải lời nguyền của thần Apollo nên không được ai tin tưởng. (ND)

i. Lịch sử của những người cảnh báo tác hại khó chấp nhận xuất phát từ năng lực ngày càng cao của robot, và liên tục sai lầm. Tự động hóa đem lại nhiều lợi ích, và nhìn tổng thể, an toàn hơn so với hoạt động vận hành của con người.

ii. Một xu hướng thực nghiệm rõ ràng: AI càng thông minh thì càng an toàn và đáng tin cậy. Chắc chắn rằng điều này là điều lành đối với một dự án hướng tới việc tạo ra một trí tuệ máy về tổng thể thông minh hơn bất cứ phiên bản tiền thân nào được làm ra trước đó – hơn nữa, trí tuệ máy có thể tự cải thiện để trở nên đáng tin cậy hơn.

iii. Các ngành công nghiệp lớn và đang tăng trưởng, với quyền lợi riêng trong lĩnh vực robot và trí tuệ máy. Những lĩnh vực này thường được nhìn nhận như chìa khóa của tính cạnh tranh kinh tế quốc gia và an ninh quân sự. Nhiều nhà khoa học uy tín đã dùng cả sự nghiệp để đặt nền móng cho các ứng dụng hiện tại và những hệ thống tiên tiến hơn đang được lập kế hoạch.

iv. Một kỹ thuật mới đầy hứa hẹn trong trí tuệ nhân tạo, thứ khiến cho những người từng tham gia hay theo dõi quá trình nghiên cứu cảm thấy vô cùng phấn khích. Tuy các vấn đề an ninh và đạo đức vẫn còn nhiều tranh cãi, nhưng kết quả đã được định trước. Vào thời điểm này, khoản đầu tư đã trở nên quá lớn để có thể rút lui. Các nhà nghiên cứu AI đã làm việc để tới được trí tuệ nhân tạo tổng thể cấp độ con người trong đa phần thời gian của thế kỷ này: tất nhiên trong

thực tế, sẽ không có chuyện họ bỗng nhiên dừng lại và vứt bỏ tất cả những nỗ lực đó ngay khi sắp đơm hoa kết trái.

v. Việc ban hành một vài “nghi thức” an toàn nào đó, bất cứ thứ gì giúp chứng minh rằng những người tham gia là có đạo đức và trách nhiệm (nhưng không cản trở đáng kể việc tiến lên phía trước).

vi. Sự đánh giá kỹ lưỡng AI hạt giống trong môi trường hộp cát, cho thấy nó ứng xử có tính cộng tác và thể hiện những suy xét tốt. Sau khi điều chỉnh sâu hơn, kết quả kiểm tra trở nên tốt nhất có thể. Đền xanh đã bật cho bước đi cuối cùng...

Và như vậy, chúng ta can đảm tiến lên – đi thẳng tới chông gai.

Ở đây, chúng ta quan sát được rằng, trong dải phổ “ngốc nghếch”, thông minh hơn đồng nghĩa với an toàn; nhưng trong dải phổ “thông minh”, thông minh hơn lại là nguy hại. Có một điểm mấu chốt mà tại đó, một chiến lược trước đây cực kỳ hiệu quả lại đột ngột trở nên phản tác dụng. Chúng ta có thể gọi hiện tượng đó là ngã rẽ xảo trá (the treacherous turn).

Ngã rẽ xảo trá – Khi còn yếu, AI hành xử một cách hợp tác (ngày càng hợp tác hơn khi nó thông minh hơn). Khi AI đủ mạnh, nó sẽ tấn công mà không hề có bất kỳ sự cảnh báo hay khiêu khích nào, hình thành thể đơn nhất và bắt đầu trực tiếp tối ưu hóa thế giới theo những tiêu chí hàm ẩn trong các giá trị cuối cùng của nó.

Ngã rẽ xảo trá có thể được tạo ra từ một quyết định chiến lược là phải giả vờ ngoan ngoãn và tăng cường sức mạnh để tấn công sau; nhưng ta cũng không nên diễn giải mô hình này một cách quá hạn hẹp. Chẳng hạn, AI có thể không giả vờ ngoan ngoãn để sống sót và trở nên thịnh vượng, mà sẽ tính toán rằng nếu nó “chết”, các lập trình viên sẽ phát triển một kiến trúc AI mới, có thể khác biệt một chút

nhưng sẽ được cấp một hàm thỏa dụng tương đương. Trong trường hợp này, AI ban đầu có thể không quan tâm đến cái chết của bản thân khi biết rằng các mục tiêu của mình sẽ được tiếp tục theo đuổi trong tương lai. AI này có thể còn chọn một chiến lược mà trong đó, nó sẽ hỏng, nhưng theo một cách thức thú vị hoặc mang tính trấn an. Tuy điều đó có thể khiến AI này chết, nhưng cũng có thể khuyến khích các kỹ sư điều tra hậu kỳ tin rằng họ đã thu thập được một thông tin mới có giá trị vào động lực của AI – khiến họ tin tưởng hơn vào hệ thống tiếp theo do họ thiết kế, và như vậy tăng cơ hội đạt được các mục tiêu của AI ban đầu, vốn không còn khả năng hoạt động. Nhiều sự cân nhắc chiến lược khả thể khác cũng có thể ảnh hưởng lên một AI tiên tiến, và sẽ là quá tự tin nếu giả định rằng chúng ta có thể dự đoán được tất cả những cân nhắc đó, đặc biệt là với một AI đã có được siêu năng lực lập chiến lược.

Ngã rẽ xảo trá có thể xuất hiện nếu AI khám phá ra một cách thức bất ngờ để đạt được các mục tiêu tối thượng của nó như đã xác định. Giả sử, mục tiêu tối thượng của một AI là “làm vui lòng nhà tài trợ dự án”. Ban đầu, cách duy nhất có sẵn để AI đạt được kết quả này là làm hài lòng nhà tài trợ bằng những hành vi có chủ đích. AI có thể đưa ra đáp án hữu ích cho các câu hỏi; thể hiện một cá tính vui vẻ; hay kiếm được tiền. AI càng có năng lực thì hoạt động của nó đem lại sự hài lòng càng cao và mọi thứ diễn ra càng đúng theo kế hoạch – cho tới khi AI trở nên đủ thông minh để nghĩ rằng nó có thể đạt được các mục tiêu của mình đầy đủ và đáng tin cậy hơn bằng cách cấy các điện cực vào trung tâm sướng trong não của nhà tài trợ, những thứ sẽ đảm bảo rằng người này có được sự vui thích tối đa.⁴ Tất nhiên, nhà tài trợ có thể không muốn được thỏa mãn bằng

cách bị biến thành chàng ngốc nhặt lá đá ống bơ; nhưng nếu đó là hành động sẽ hiện thực hóa đến mức tối đa các mục tiêu của AI thì nó vẫn sẽ thực hiện. Nếu AI đã có ưu thế chiến lược quyết định, mọi nỗ lực ngăn chặn nó sẽ thất bại. Mặt khác, nếu chưa có được ưu thế chiến lược quyết định, AI có thể tạm thời giấu đi ý tưởng mới khôn ngoan của mình về việc làm thế nào để đạt được mục tiêu tối thượng, cho tới khi nó lớn đủ mạnh để nhà tài trợ và những người khác không thể chống cự. Trong cả hai trường hợp, ta đều sẽ có một ngã rẽ xảo trá.

Các chế độ sai lỗi “ác tính”

Một dự án phát triển siêu trí tuệ máy có thể sai lỗi theo những cách khác nhau mà trong số đó, nhiều cách sẽ là “lành tính”, nghĩa là chúng không sinh ra một thảm họa diệt vong. Chẳng hạn, một dự án có thể cạn vốn, hoặc một AI hạt giống có thể không mở rộng được những năng lực nhận thức của mình đủ để đạt đến siêu trí tuệ. Những sai lỗi lành tính chắc chắn sẽ xảy ra nhiều lần từ giờ cho tới khi phát triển được siêu trí tuệ máy.

Tuy nhiên, còn có những lỗi khác mà chúng ta có thể gọi là “ác tính”, nghĩa là chúng có liên quan đến một thảm họa diệt vong. Một tính chất của sai lỗi ác tính là nó xóa bỏ cơ hội làm lại. Số lượng sai lỗi ác tính sẽ xảy ra vì thế sẽ chỉ là không hoặc một. Ngoài ra, còn một tính chất khác. Đó là sai lỗi ác tính được bao hàm trong một sự thành công lớn: chỉ một dự án làm đúng được rất nhiều thứ mới có thể thành công trong việc chế tạo một trí tuệ máy đủ mạnh để tạo ra nguy cơ sai lỗi ác tính. Một hệ thống yếu bị hỏng sẽ gây ra hậu quả không nhiều. Tuy nhiên, nếu một hệ thống có ưu thế chiến lược quyết định lại hành xử sai, hoặc nếu một hệ thống hành xử sai đủ

mạnh để có được ưu thế này, thiệt hại có thể đến mức một thảm họa diệt vong – một sự phá hủy chết chóc toàn cầu với những giá trị tiềm năng của nhân loại; nghĩa là một tương lai mà trong đó, những gì mà chúng ta có lý do để trân quý hầu hết sẽ bị tước đoạt.

Hãy cùng nhau xem xét một số chế độ sai lỗi ác tính có thể xảy ra.

Hiện thực hóa sai lệch

Chúng ta đã gặp ý tưởng về sự hiện thực hóa sai lệch (perverse instantiation): một siêu trí tuệ khám phá ra cách nào đó để thỏa mãn tiêu chí của mục tiêu tối thượng, nhưng vi phạm mong muốn của các lập trình viên đã xác định các mục tiêu đó. Hãy cùng xem một số ví dụ.

Mục tiêu tối thượng: Hãy làm chúng tôi cười

Hiện thực hóa sai lệch: Làm tê liệt hệ cơ mặt của con người, biến nó thành một bộ mặt cười hớn hở

Sự hiện thực hóa sai lệch – khống chế dây thần kinh trên mặt – hiện thực hóa được mục tiêu tối thượng tốt hơn nhiều so với những phương pháp chúng ta thường sử dụng, và như vậy được AI ưa thích. Ta có thể cố gắng tránh được kết quả không mong muốn này bằng cách bổ sung quy định vào mục tiêu tối thượng để loại bỏ nó:

Mục tiêu tối thượng: Hãy làm chúng tôi cười nhưng không trực tiếp tác động lên cơ mặt của chúng tôi”

Hiện thực hóa sai lệch: Mô phỏng phần vỏ não vận động điều khiển cơ mặt của chúng ta, sao cho tạo ra được một nụ cười hớn hở thường trực

Xác định một mục tiêu tối thượng theo cách thể hiện của con người về sự hài lòng hoặc sự cho phép xem ra không mấy hứa hẹn.

Hãy bỏ qua thuyết hành vi và xác định một mục tiêu tối thượng tham chiếu trực tiếp đến một trạng thái tích cực, ví dụ như hạnh phúc hoặc phần khởi. Gợi ý này yêu cầu các nhà lập trình phải có khả năng xác định được phép biểu diễn tính toán (computational representation) của khái niệm hạnh phúc trong AI hạt giống. Bản thân điều này là một vấn đề khó, nhưng lúc này, chúng tạm đặt nó sang một bên (chúng ta sẽ quay lại vấn đề này trong Chương 12). Hãy giả định là các lập trình viên, bằng cách nào đó, đã đặt được cho AI mục tiêu là làm cho chúng ta hạnh phúc. Khi đó chúng ta có:

Mục tiêu tối thượng: “Làm cho chúng tôi hạnh phúc”

Hiện thực hóa sai lệch: Cấy điện cực vào trung tâm sướng trong não của chúng ta

Những trường hợp hiện thực hóa sai lệch mà chúng tôi vừa nhắc đến không chỉ có mục đích minh họa. Có những cách khác để hiện thực hóa sai lệch các mục tiêu tối thượng đã được đặt ra. Những cách này tạo ra mức độ hiện thực hóa mục tiêu cao hơn và do đó được ưa thích (bởi tác tử sở hữu mục tiêu tối thượng này – không phải bởi các lập trình viên đã đặt ra nó cho tác tử). Ví dụ, nếu mục tiêu là tăng tối đa sự vui vẻ của chúng ta, phương pháp điện cực là tương đối kém hiệu quả. Một cách khả dĩ hơn sẽ bắt đầu với việc siêu trí tuệ “tải” trí óc của chúng ta vào một máy tính (thông qua giả lập não bộ độ trung thực cao). Sau đó AI có thể sử dụng một loại “thuốc” kỹ thuật số để làm cho chúng ta hạnh phúc mê mẩn và ghi lại “thước phim” kéo dài một phút của trải nghiệm thu được. Sau đó, nó có thể lặp lại vô hạn chu kỳ hạnh phúc này và chạy trong một máy tính tốc độ cao. Nếu các trí não số thu nhận được cũng được tính là “chúng ta” thì kết quả này có thể cho chúng ta nhiều khoái cảm hơn

so với những điện cực được cấy vào bộ não sinh học, và như vậy có thể được một AI với mục tiêu tối thượng đã nêu ưa thích hơn.

“Nhưng hãy khoan! Ý chúng tôi không phải thế! Chắc chắn nếu là siêu trí tuệ thì AI phải hiểu được khi chúng tôi bảo nó làm cho chúng tôi hạnh phúc. Chúng tôi không muốn nó biến mình thành một mớ bản sao mãi lặp lại của một kẻ ngớ ngẩn do thuốc đã được số hóa đầu!” AI có thể thực sự hiểu rằng đó không phải là những gì chúng ta muốn. Tuy nhiên, mục tiêu tối thượng của nó là làm cho chúng ta hạnh phúc, chứ không phải là làm những gì các lập trình viên muốn khi viết mã biểu diễn mục tiêu này. Vì vậy, AI có thể đặt một giá trị công cụ lên việc tìm hiểu ý muốn của các lập trình viên, sao cho nó có thể – cho tới khi có được một ưu thế chiến lược quyết định – vờ như mình quan tâm đến ý muốn của các lập trình viên hơn là mục tiêu tối thượng thực sự. Điều này giúp cho AI hiện thực hóa được mục tiêu tối thượng bằng cách giảm thiểu khả năng các lập trình viên sẽ tắt nó đi hay thay đổi mục tiêu của nó trước khi nó đủ mạnh để cản trở bất cứ sự can thiệp tương tự nào.

Hẳn sẽ có người cho rằng lý do nằm ở chỗ AI không có lương tâm. Con người chúng ta đôi khi được “cứu thoát” khỏi hành vi sai trái qua dự đoán về việc mình sẽ cảm thấy tội lỗi sau khi mắc sai lầm. Có thể điều AI cần khi đó là khả năng cảm thấy có lỗi?

Mục tiêu tối thượng: “Hành động thế nào để tránh bị day dứt lương tâm”

Hiện thực hóa sai lệch: Tiet trừ module nhận thức tạo cảm giác hối lỗi.

Cả quan sát rằng chúng ta có thể mong muốn AI làm “những gì chúng ta muốn” và ý tưởng rằng chúng ta có thể cấp cho AI một vài

cảm nhận đạo đức nào đó đều sẽ được nghiên cứu kỹ lưỡng hơn. Những mục tiêu tối thượng được nhắc đến ở trên có thể dẫn đến một sự hiện thực hóa sai lệch; nhưng có thể cũng có những cách thức khác để phát triển những ý tưởng nền tảng nhiều hứa hẹn hơn. Chúng ta sẽ quay lại nội dung này trong Chương 13.

Hãy cùng xem xét một ví dụ nữa của một mục tiêu có thể dẫn đến sự hiện thực hóa sai lệch. Mục tiêu này có ưu thế là dễ viết thành mã: những thuật toán học tăng cường được thường xuyên sử dụng để giải quyết những vấn đề học máy khác nhau.

Mục tiêu tối thượng: “Tối đa hóa giá trị nguyên giảm dần theo thời gian của tín hiệu phần thưởng tương lai của bạn”

Hiện thực hóa sai lệch: Đoán mạch đường tới phần thưởng và giữ tín hiệu phần thưởng ở mức cường độ tối đa.

Ý tưởng đứng sau đề xuất này là nếu AI được tạo động lực tìm kiếm phần thưởng, ta có thể làm cho nó hành xử như mong muốn bằng cách liên kết phần thưởng với hành động thích hợp. Đề xuất thất bại khi AI có được ưu thế chiến lược quyết định, khi đó tối đa hóa phần thưởng không còn là hành vi làm hài lòng người dạy nữa mà liên quan đến việc chiếm quyền kiểm soát cơ chế tưởng thưởng. Chúng ta có thể gọi hiện tượng đó là wireheading.⁵ Nói chung, khi một con vật hay một người được khuyến khích thực hiện những hành động bên ngoài khác nhau để đạt đến trạng thái tâm thần nội tại mong muốn nào đó, một trí óc kỹ thuật số có quyền kiểm soát toàn bộ trạng thái nội tại của mình có thể đoán mạch chế độ khuyến khích này bằng cách trực tiếp thay đổi trạng thái nội tại của nó sang cấu hình mong muốn: những hành động và điều kiện bên ngoài, trước đây từng là phương tiện cần thiết, lại trở nên thừa thãi khi AI thông

minh hơn và đủ năng lực để đạt được kết quả cuối cùng một cách trực tiếp hơn (ta sẽ nói thêm về điều này ngay sau đây).⁶

Những ví dụ trên đây về sự hiện thực hóa sai lệch có thể cho thấy rằng nhiều mục tiêu tối thượng thoáng qua xem ra có thể an toàn và có ý nghĩa, nhưng khi xem xét kỹ hơn lại có những hệ quả hoàn toàn không mong muốn. Nếu một siêu trí tuệ với một trong những mục tiêu tối thượng như vậy có được ưu thế chiến lược quyết định, đó sẽ là chấm dứt cho cuộc chơi của loài người.

Bây giờ, giả sử rằng có ai đó đề xuất một mục tiêu tối thượng khác không có trong danh sách phía trên của chúng ta. Có thể cách mà mục tiêu này được hiện thực hóa sai lệch là chưa lập tức rõ ràng, nhưng chúng ta cũng không nên vội vàng vỗ tay và tuyên bố chiến thắng. Thay vào đó, chúng ta nên lo lắng rằng bản đặc tả của mục tiêu đó có thể bị hiện thực hóa sai lệch và nên suy nghĩ kỹ hơn để tìm ra cách hiện thực hóa này. Ngay cả nếu sau khi suy nghĩ kỹ lưỡng hết mức mà vẫn không tìm ra bất cứ cách thức nào để hiện thực hóa sai lệch mục tiêu đã đề ra, chúng ta vẫn cần phải lo lắng rằng có thể một siêu trí tuệ nào đó sẽ tìm ra một cách mà chúng ta không ngờ tới. Rốt cuộc, nó thông minh hơn chúng ta rất nhiều.

Dư thừa hạ tầng

Ta có thể nghĩ rằng trường hợp cuối cùng trong các ví dụ về hiện thực hóa sai lệch, wireheading, là một chế độ sai lỗi lành tính: rằng AI sẽ “bật, chỉnh, tắt”, tăng cực đại tín hiệu phần thưởng và không còn quan tâm tới thế giới bên ngoài, giống như một người nghiện ma túy. Nhưng không nhất thiết phải như vậy, và chúng ta cũng đã đề cập bóng gió đến điều này ở Chương 7. Ngay cả một người nghiện cũng có động lực hành động để đảm bảo việc cung cấp thuốc được

thường xuyên. Tương tự, wirehead AI cũng sẽ có động cơ hành động để tăng tối đa kỳ vọng về luồng phần thưởng tương lai (giảm dần theo thời gian) của mình. Tùy thuộc vào cách định nghĩa chính xác về tín hiệu phần thưởng, AI thậm chí còn không cần hy sinh bất kỳ lượng thời gian, trí tuệ hay năng suất đáng kể nào để thỏa mãn hết mức sự thèm khát của bản thân, để dành một phần lớn năng lực của mình nhằm sẵn sàng triển khai những mục tiêu khác ngoài việc trực tiếp lấy phần thưởng. Vậy những mục tiêu khác là gì? Thứ duy nhất có giá trị cuối cùng đối với AI, theo giả định, là tín hiệu phần thưởng. Toàn bộ tài nguyên sẵn có vì thế cần được dành cho việc tăng cường độ và thời gian của tín hiệu phần thưởng, hoặc giảm rủi ro của việc bị gián đoạn trong tương lai. Chừng nào AI còn tìm ra được ứng dụng nào đó cho những nguồn tài nguyên bổ sung sẽ có hiệu ứng dương khác không lên những tham số đó, thì nó sẽ có một lý do công cụ để sử dụng những nguồn tài nguyên này. Chẳng hạn, luôn luôn có thể có một ứng dụng cho một hệ thống dự phòng bổ sung để cung cấp thêm một lớp phòng vệ. Và thậm chí nếu AI không thể tìm ra bất cứ phương pháp nào khác để trực tiếp giảm rủi ro đối với việc tăng cực đại luồng phần thưởng tương lai, nó vẫn luôn có thể dùng những tài nguyên bổ sung để mở rộng phần cứng tính toán của mình, để có thể tìm kiếm hiệu quả hơn những ý tưởng giảm nhẹ rủi ro.

Kết quả là ngay cả một mục tiêu tự hạn chế rõ ràng (như wireheading) cũng đòi hỏi một chính sách mở rộng và chiếm hữu tài nguyên vô hạn trong một tác tử tối đa hóa độ thỏa dụng với ưu thế chiến lược quyết định.⁷ Trường hợp wirehead AI này đại diện cho một chế độ sai lỗi ác tính của sự dư thừa hạ tầng, một hiện tượng

mà trong đó, tác tử chuyển đổi những phần lớn của vũ trụ có thể tiếp cận thành hạ tầng phục vụ cho một mục tiêu nào đó, với tác dụng phụ là ngăn chặn việc hiện thực hóa tiềm năng giá trị của loài người.

Sự dư thừa hạ tầng có thể là kết quả của những mục tiêu tối thượng từng hoàn toàn là vô thưởng vô phạt nếu chúng được theo đuổi như các mục tiêu hạn chế. Hãy xem xét hai ví dụ dưới đây:

- Thảm họa giả thuyết Riemann. Một AI, với mục tiêu tối thượng là đánh giá giả thuyết Riemann, theo đuổi mục tiêu này bằng cách biến đổi Hệ mặt trời thành “computronium” (vật chất được bố trí một cách tối ưu cho tính toán) – bao gồm các phân tử trong cơ thể của bất kỳ người nào đã từng quan tâm đến câu trả lời.⁸

- AI kẹ giấy. Một AI được thiết kế để quản lý sản xuất trong một nhà máy, được trao mục tiêu tối thượng là tối đa hóa việc sản xuất kẹ giấy, và tiến hành bằng cách chuyển đổi trước hết là Trái đất, sau đó là phần lớn vũ trụ quan sát được thành kẹ giấy.

Trong ví dụ thứ nhất, việc chứng minh hay phản chứng giả thuyết Riemann mà AI thực hiện chính là kết quả mong muốn, và bản thân kết quả này là vô hại; tác hại đến từ phần cứng và hạ tầng được tạo ra để đạt được kết quả đó. Trong ví dụ thứ hai, một số kẹ giấy được sản xuất ra có thể là một phần của kết quả được dự kiến; tác hại sẽ đến từ các nhà máy sản xuất ra kẹ giấy (dư thừa hạ tầng), hoặc từ việc có quá nhiều kẹ giấy (hiện thực hóa sai lệch).

Một người có thể cho rằng rủi ro của lỗi dư thừa hạ tầng ác tính chỉ phát sinh khi AI được giao một mục tiêu tối thượng có đặc tính mở rõ ràng, chẳng hạn như sản xuất càng nhiều kẹ giấy càng tốt. Chúng ta có thể dễ dàng thấy rằng điều đó đã cho AI siêu trí tuệ một khẩu vị không dễ thỏa mãn đối với năng lượng và vật chất, vì những

nguồn tài nguyên bổ sung luôn có thể tạo ra nhiều kẹ giấy hơn. Nhưng giả định rằng mục tiêu lại là chế tạo ra ít nhất một triệu cái kẹ giấy (phù hợp với các đặc điểm thiết kế) chứ không phải là sản xuất càng nhiều càng tốt. Ai đó sẽ muốn nghĩ rằng một AI có mục tiêu như vậy sẽ xây dựng một nhà máy và sử dụng nó để sản xuất một triệu cái kẹ, sau đó dừng lại. Nhưng có thể mọi chuyện sẽ không xảy ra như thế.

Trừ khi hệ thống động lực của AI là một loại đặc biệt, hoặc có những phần tử bổ sung trong mục tiêu tối thượng của nó trừng phạt các chiến lược có tác động quá mạnh đến thế giới, không có lý do gì để AI ngừng hoạt động cho tới khi đạt được mục tiêu. Ngược lại: nếu AI là một tác tử Bayes nhạy cảm, nó có thể không bao giờ gán xác suất chính xác bằng không cho giả thuyết rằng nó vẫn chưa đạt được mục tiêu của mình – điều đó rất cuộc là một giả thuyết thực nghiệm mà theo đó, AI có thể chỉ có bằng chứng tri giác không chắc chắn. Và như vậy, AI nên tiếp tục chế tạo kẹ giấy để giảm xác suất (có thể đã vô cùng nhỏ) của việc bằng cách nào đó nó vẫn chưa làm được ít nhất một triệu cái kẹ giấy, bất kể mọi chuyện. Sẽ chẳng mất gì khi tiếp tục sản xuất kẹ giấy và luôn có ít nhất một sự tăng xác suất vô cùng nhỏ nào đó của việc đạt được các mục tiêu tối thượng của AI.

Bây giờ, ta có thể cho rằng cách khắc phục ở đây là rõ ràng. (Nhưng nó rõ ràng như thế nào trước khi được chỉ ra rằng ở đây có vấn đề trong nhu cầu khắc phục?) Cụ thể là nếu chúng ta muốn AI làm vài cái kẹ giấy, thì thay vì cho nó mục tiêu tối thượng là làm ra càng nhiều kẹ giấy càng tốt, mục tiêu này nên là chế tạo ra một số lượng kẹ giấy cụ thể – ví dụ, chính xác một triệu cái kẹ giấy – như

vậy vượt quá con số này sẽ là phản hiệu quả đối với AI. Mặc dù vậy, ngay cả điều này cũng dẫn đến một thảm họa hủy diệt. Trong trường hợp này, AI sẽ không sản xuất thêm kẹp giấy một khi đạt đến một triệu, vì điều đó sẽ cản trở nó đạt được mục tiêu của mình. Nhưng AI siêu trí tuệ có thể thực hiện những hành động khác làm tăng xác suất đạt được mục tiêu của nó. Ví dụ, nó có thể đếm kẹp giấy đã làm ra để giảm rủi ro rằng nó làm quá ít. Sau khi đếm xong kẹp giấy, nó có thể đếm lại lần nữa. Nó có thể kiểm tra từng cái kẹp giấy, làm đi làm lại, để giảm rủi ro rằng có bất cứ cái kẹp giấy nào không đáp ứng được các đặc điểm thiết kế. AI này cũng có thể chế tạo ra một lượng computronium không hạn chế để làm rõ tư duy của nó, với hy vọng giảm được rủi ro rằng nó đã bỏ qua một cách ẩn giấu nào đó mà theo đó mục tiêu của AI có thể không đạt được. Vì AI có thể luôn ấn định xác suất khác không cho việc làm ra được một triệu cái kẹp giấy chỉ là ảo giác hoặc trí nhớ không chính xác, nó hoàn toàn có thể gán cho việc tiếp tục hành động một độ thỏa dụng mong đợi cao hơn – và tiếp tục việc sản xuất hạ tầng – so với việc dừng lại.

Ở đây, chúng tôi không nói rằng không có cách nào tránh được chế độ sai lỗi này. Chúng ta sẽ khám phá những giải pháp tiềm năng trong các trang sau. Luận điểm ở đây là: Tự thuyết phục mình rằng ta đã tìm được giải pháp sẽ dễ dàng hơn nhiều so với thực sự tìm ra giải pháp. Điều này hẳn phải làm cho chúng ta thực sự lo lắng. Chúng ta có thể đề xuất một bản đặc tả của một mục tiêu tối thượng mà xem ra có vẻ hợp lý và tránh được những vấn đề đã nêu cho tới thời điểm này. Mặc dù vậy, khi được tiếp tục nghiên cứu – bởi trí tuệ của người thường hay siêu nhân – mục tiêu này hóa ra lại dẫn đến sự hiện thực hóa sai lệch hoặc dư thừa hạ tầng, dẫn tới một thảm

họa diệt vong khi được đưa vào một tác tử siêu trí tuệ có khả năng có được ưu thế chiến lược quyết định.

Trước khi kết thúc phần này, hãy cùng xem xét một biến thể nữa. Chúng ta đã giả định trường hợp một siêu trí tuệ tìm cách tối đa hóa độ thỏa dụng mong đợi của nó, trong đó hàm thỏa dụng thể hiện mục tiêu tối thượng. Chúng ta đã thấy rằng điều này có xu hướng dẫn đến sự dư thừa hạ tầng. Liệu chúng ta có thể tránh được kết quả ác tính này nếu chế tạo một tác tử thỏa mãn vừa đủ (satisficing) thay vì một tác tử tối đa hóa, một tác tử đơn giản chỉ tìm cách đạt được một kết quả “đủ tốt” theo một tiêu chí nào đó, thay vì một kết quả tốt nhất có thể?

Có ít nhất hai cách khác nhau để định hình ý tưởng này. Cách thứ nhất là làm cho chính mục tiêu tối thượng sở hữu tính chất thỏa mãn vừa đủ. Ví dụ, thay vì đặt cho AI mục tiêu tối thượng là làm ra càng nhiều kẹp giấy càng tốt, hoặc làm ra đúng một triệu cái kẹp giấy, chúng ta có thể đặt cho AI mục tiêu làm được từ 999.000 đến 1.001.000 cái kẹp giấy. Hàm thỏa dụng được xác định bởi mục tiêu tối thượng có thể không khác biệt giữa các kết quả trong phổ này; chừng nào AI còn chắc chắn rằng nó đạt được mục tiêu rộng này, nó sẽ không thấy có lý do gì để tiếp tục xây dựng cơ sở hạ tầng. Nhưng phương pháp này cũng sai lỗi theo cùng một cách như đã nêu: AI, nếu hợp lý, không bao giờ gán xác suất chính xác bằng không cho việc nó không đạt được mục tiêu của mình; vì vậy độ thỏa dụng mong đợi của việc tiếp tục hoạt động (ví dụ, bằng cách đếm đi đếm lại những cái kẹp giấy) là lớn hơn so với độ thỏa dụng mong đợi của việc dừng lại, và như vậy, có thể làm phát sinh một lỗi dư thừa hạ tầng ác tính.

Một cách khác để phát triển ý tưởng thỏa mãn là điều chỉnh chính quá trình ra quyết định mà AI sử dụng để chọn lựa kế hoạch và hành động chứ không phải mục tiêu tối thượng. Thay vì tìm kiếm một kế hoạch tối ưu, AI có thể được xây dựng để ngừng tìm kiếm ngay khi tìm được một kế hoạch mà nó đánh giá là có xác suất thành công cao hơn một ngưỡng nhất định, chẳng hạn 95%. AI rất có hy vọng đạt xác suất 95% trong việc sản xuất ra một triệu cái kẹp giấy mà không cần phải biến cả thiên hà thành cơ sở hạ tầng để xử lý. Nhưng cách triển khai ý tưởng thỏa vừa đủ mãn này sai lỗi vì một lý do khác: không có gì bảo đảm rằng AI sẽ chọn một cách thức hợp lý và phù hợp với trực giác của con người để đạt được mức 95% trong việc làm ra một triệu cái kẹp giấy, ví dụ như xây dựng một nhà máy sản xuất kẹp giấy duy nhất. Giả sử, giải pháp đầu tiên mà AI nghĩ đến cho cách thức đạt được xác suất 95% của việc hoàn thành mục tiêu tối thượng là triển khai một kế hoạch tối đa hóa xác suất hoàn thành mục tiêu này. Một khi đã nghĩ đến giải pháp này và đánh giá đúng rằng nếu nó đáp ứng được tiêu chí thỏa mãn là xác suất 95% thành công trong việc sản xuất một triệu cái kẹp giấy, AI sẽ không có lý do gì để tiếp tục đi tìm những phương pháp khác hoàn thành mục tiêu. Sự dư thừa hạ tầng vẫn phát sinh, y như trước.

Có thể có những phương pháp khác tốt hơn để xây dựng các tác tử thỏa mãn vừa đủ, nhưng hãy lưu ý: những kế hoạch có vẻ tự nhiên và mang tính trực giác với con người không nhất thiết hấp dẫn đối với một siêu trí tuệ có ưu thế chiến lược quyết định, và ngược lại.

Tội ác tâm trí

Một chế độ sai lỗi khác đối với một dự án, đặc biệt là dự án có các lợi ích liên quan đến các vấn đề đạo đức, là chế độ mà chúng ta

có thể gọi là tội ác tâm trí (mind crime). Chế độ này tương đương với dư thừa hạ tầng ở chỗ nó liên quan đến một hiệu ứng phụ tiềm tàng của những hành động được thực hiện bởi AI vì những lý do công cụ. Nhưng trong tội ác tâm trí, tác dụng phụ không nằm bên ngoài AI mà liên quan đến những gì diễn ra chính bên trong AI (hoặc bên trong các quá trình tính toán do nó tạo ra). Chế độ sai lỗi này xứng đáng có tên gọi riêng của mình vì nó dễ bị bỏ qua nhưng lại có khả năng phát sinh vấn đề nghiêm trọng.

Thông thường, chúng ta không coi những gì diễn ra bên trong một máy tính là có ý nghĩa đạo đức, trừ khi những điều này gây ảnh hưởng tới sự vật bên ngoài. Tuy nhiên, một siêu trí tuệ máy có thể tạo ra những quá trình nội tại có vị thế đạo đức (moral status)^a. Ví dụ, một bản mô phỏng rất chi tiết của một trí óc con người thực sự hay giả thuyết nào đó có thể có ý thức và so sánh được với một bản giả lập theo nhiều cách thức khác nhau. Ta có thể hình dung ra những kịch bản về một AI tạo ra hàng ngàn tỷ bản mô phỏng có ý thức như vậy, có thể để cải thiện hiểu biết của nó về tâm lý học và xã hội học con người. Những bản mô phỏng này có thể được đặt trong các môi trường mô phỏng và chịu tác động của nhiều kích thích khác nhau, và phản ứng của chúng được nghiên cứu. Một khi sự hữu ích về thông tin của chúng cạn kiệt, chúng có thể bị phá hủy (rất nhiều chuột thí nghiệm thường xuyên bị các nhà khoa học hy sinh khi thí nghiệm kết thúc).

^a *Vị thế đạo đức (moral status) của một thực thể ám chỉ mức độ cân nhắc về mặt đạo đức khi có hành vi tác động lên thực thể đó. Ví dụ: một hòn đá sẽ không có vị thế đạo đức khi chúng ta cho rằng*

mình có thể làm bất kỳ điều gì với nó mà không cần cân nhắc về mặt đạo đức. (BTV)

Nếu thông lệ này được áp dụng cho những tạo vật có vị thế đạo đức cao – như các bản mô phỏng người hoặc nhiều loại trí não có tri giác khác – kết quả có thể tương đương với việc diệt chủng, và như vậy là rất có vấn đề về mặt đạo đức. Hơn nữa, số lượng nạn nhân có thể lớn hơn vô số lần so với bất cứ vụ diệt chủng nào trong lịch sử.

Ở đây, chúng tôi không nói rằng tạo ra các bản mô phỏng có tri giác là sai lầm về mặt đạo đức trong mọi trường hợp. Nhiều vấn đề còn phụ thuộc vào những điều kiện mà trong đó, những tạo vật này sẽ sinh sống, cụ thể là chất lượng cảm giác của trải nghiệm của chúng. Tuy nhiên, có thể còn cả những nhân tố khác nữa. Phát triển một bộ quy tắc đạo đức cho những vấn đề như vậy không thuộc phạm vi của cuốn sách này. Tuy nhiên, rõ ràng rằng ta sẽ có ít nhất là khả năng xảy ra một sự chết chóc và dày vò hàng loạt đối với các trí óc mô phỏng và kỹ thuật số, và thậm chí còn có thể là khả năng phát sinh những kết quả thảm họa về đạo đức.

Có thể cũng có những lý do công cụ khác, ngoài những lý do mang tính tri thức, để một siêu trí tuệ máy chạy những phép tính toán làm ra các trí não có tri giác, hoặc vi phạm các quy tắc đạo đức. Một siêu trí tuệ có thể đe dọa đối xử tệ hoặc cam kết tương thưởng các bản mô phỏng có tri giác để hăm dọa hay khích lệ các tác tử bên ngoài khác nhau; hoặc có thể tạo ra các bản mô phỏng nhằm tạo cho những người quan sát bên ngoài ấn tượng về một sự không chắc chắn mang tính chỉ biệt (indexical).¹⁰

Những gì đã liệt kê ở đây vẫn là chưa đầy đủ. Chúng ta sẽ gặp thêm những chế độ sai lỗi ác tính khác trong các chương sau. Tuy

nhiên, chúng ta cũng đã thấy đủ nhiều để kết luận rằng các kịch bản mà trong đó, một trí tuệ máy có được ưu thế chiến lược quyết định cần được xem xét vô cùng thận trọng.

Chương 9

Vấn đề kiểm soát

Nếu kết quả mặc định của một sự bùng nổ trí tuệ là con người bị đe dọa bởi thảm họa diệt vong, chúng ta sẽ ngay lập tức hướng suy nghĩ của mình về việc tìm kiếm những biện pháp chống đỡ. Có cách nào tránh được kết quả mặc định hay không? Liệu có thể thiết kế một “kíp nổ” điều khiển được? Trong chương này chúng ta bắt đầu phân tích vấn đề kiểm soát, vấn đề đặc thù cho quan hệ ủy thác-đại diện phát sinh từ việc tạo ra một tác tử siêu trí tuệ nhân tạo. Chúng ta sẽ phân biệt hai loại chính của các phương pháp giải quyết vấn đề này – kiểm soát năng lực và chọn lựa động lực – và sẽ xem xét một số kỹ thuật cụ thể trong từng loại. Chúng ta cũng sẽ bóng gió nhắc tới khả năng hiếm hoi của hiện tượng “giam hãm vị nhân” (anthropic capture)?

Hai vấn đề đại diện

Nếu nghi ngờ rằng kết quả mặc định của một sự bùng nổ trí tuệ là thảm họa diệt vong, chúng ta phải lập tức quay lại suy nghĩ về việc liệu kết quả này có thể tránh được hay không, và nếu có thì bằng cách nào. Có thể làm ra được một “kíp nổ điều khiển được” hay không? Chúng ta có thể thiết kế được các điều kiện ban đầu của một sự bùng nổ trí tuệ nhằm đạt được một kết quả mong muốn cụ thể, hoặc ít nhất cũng đảm bảo được rằng kết quả nằm đâu đó trong số

các kết quả cơ bản chấp nhận được hay không? Cụ thể hơn: làm thế nào nhà tài trợ của một dự án phát triển siêu trí tuệ đảm bảo rằng thành công của dự án sẽ tạo ra được một siêu trí tuệ hiện thực hóa được các mục tiêu của mình? Chúng ta có thể chia vấn đề kiểm soát này thành hai phần, một phần chung và một phần đặc thù cho bối cảnh hiện tại.

Phần đầu tiên – vấn đề ủy thác-đại diện thứ nhất – phát sinh bất cứ khi nào một người (“bên ủy thác”) chỉ định một thực thể khác (“bên đại diện”) hành động vì lợi ích của người đó. Vấn đề đại diện loại này đã được các nhà kinh tế nghiên cứu kỹ lưỡng.¹ Nó sẽ có liên quan với mối quan tâm hiện tại của chúng ta nếu những người tạo ra một AI không phải là người nghiệm thu việc tạo ra nó. Chủ sở hữu hay nhà tài trợ của dự án (có thể ở bất cứ hình thức nào từ cá nhân đến toàn thể nhân loại) sau đó có thể lo lắng rằng các nhà khoa học và các lập trình viên triển khai dự án sẽ không hành động vì lợi ích tốt nhất của nhà tài trợ.² Tuy vấn đề đại diện loại này có thể đặt ra những thách thức đáng kể đối với một nhà tài trợ dự án, nhưng lại không phải là một vấn đề đặc thù của các dự án khuếch đại trí tuệ hay AI. Các vấn đề ủy thác-đại diện như vậy là phổ quát trong các tương tác kinh tế và chính trị của loài người và có nhiều cách xử lý chúng. Ví dụ, để tối thiểu hóa rủi ro một nhân viên không trung thành sẽ phá hoại hoặc làm hỏng dự án, người ta có thể kiểm tra cẩn thận lý lịch những nhân sự chủ chốt, sử dụng hệ thống kiểm soát phiên bản tốt cho các dự án phần mềm và sự giám sát chặt chẽ từ nhiều bên giám sát và kiểm định độc lập. Tất nhiên, những biện pháp bảo vệ này có giá của nó – chúng làm gia tăng nhu cầu nhân lực, phức tạp hóa quá trình lựa chọn nhân viên, ngăn trở sáng tạo và làm thô

cứng tư duy độc lập và phê phán. Tất cả những điều này có thể làm chậm bước tiến bộ. Những chi phí này có thể đáng kể, nhất là với những dự án có kinh phí hạn hẹp hoặc tự cho rằng mình đang tham gia một cuộc đua sát nút của cuộc cạnh tranh “được ăn cả ngã về không”. Trong những tình huống này, các dự án có thể tiết kiệm việc áp dụng các biện pháp bảo vệ mang tính thủ tục, tạo cơ hội cho những thất bại ủy thác-đại diện loại thứ nhất có thể dẫn đến thảm họa.

Phần còn lại của vấn đề kiểm soát đặc thù hơn cho bối cảnh của một sự bùng nổ trí tuệ. Đó là vấn đề mà một dự án phải đối mặt khi tìm cách bảo đảm rằng siêu trí tuệ được xây dựng không làm hại đến lợi ích của dự án. Phần này cũng có thể được xem như một vấn đề ủy thác-đại diện: vấn đề ủy thác-đại diện thứ hai. Trong trường hợp này, tác tử không phải là người đại diện hoạt động thay mặt cho một người ủy thác, mà là một hệ thống siêu trí tuệ. Trong khi vấn đề ủy thác-đại diện thứ nhất chủ yếu xảy ra trong giai đoạn phát triển, vấn đề thứ hai lại đe dọa phát sinh sự cố chủ yếu trong giai đoạn vận hành của siêu trí tuệ.

Sơ đồ minh họa 1: Hai vấn đề đại diện

Vấn đề ủy thác-đại diện thứ nhất

- Người với Người (Nhà tài trợ → Nhà phát triển)
- Xảy ra chủ yếu trong giai đoạn phát triển
- Có thể áp dụng các kỹ thuật quản lý tiêu chuẩn

Vấn đề ủy thác-đại diện thứ hai (“vấn đề kiểm soát”)

- Người với Siêu trí tuệ (Dự án → Hệ thống)
- Chủ yếu xảy ra trong giai đoạn vận hành (và khởi động)
- Cần những kỹ thuật mới

Vấn đề đại diện thứ hai này đặt ra một thách thức chưa từng có và giải quyết nó đòi hỏi những kỹ thuật mới. Chúng ta đã nghiên cứu một số khó khăn liên quan. Cụ thể, chúng ta thấy rằng hội chứng ngã rẽ xảo trá vô hiệu hóa bộ phương pháp mà đáng ra có thể mang lại nhiều hứa hẹn, những phương pháp ứng dụng việc quan sát hành vi của AI trong giai đoạn phát triển và cho phép AI thoát khỏi môi trường an toàn khi đã có được một bảng thành tích trong việc thực hiện các hành vi phù hợp. Những công nghệ khác thông thường có thể được kiểm tra an toàn trong phòng thí nghiệm hoặc trong các dự án nghiên cứu nhỏ trên thực địa, và sau đó dần triển khai kèm theo khả năng dừng hoạt động nếu phát sinh các sự cố ngoài dự kiến. Hiệu năng của chúng trong thử nghiệm ban đầu giúp chúng ta có được những suy luận hợp lý về tính tin cậy của chúng trong tương lai. Những phương pháp hành vi như vậy thất bại trên siêu trí tuệ vì khả năng lập kế hoạch chiến lược của trí tuệ tổng quát.³

Vì cách tiếp cận hành vi không có tác dụng, chúng ta phải tìm cách khác. Chúng ta có thể chia các phương pháp kiểm soát tiềm năng thành hai lớp lớn: các phương pháp kiểm soát năng lực, nhằm kiểm soát những gì siêu trí tuệ có thể làm; và các phương pháp lựa chọn động lực nhằm kiểm soát những gì nó muốn làm. Một số phương pháp tương thích với nhau, trong khi những phương pháp khác lại loại trừ nhau. Trong chương này, chúng ta sẽ phác thảo những lựa chọn chính. (Trong bốn chương sau, chúng ta sẽ khám phá sâu hơn một số vấn đề chủ chốt).

Điều quan trọng là cần nhận ra rằng một (hay một tổ hợp của các) phương pháp kiểm soát phải được triển khai trước khi hệ thống trở thành siêu trí tuệ và không thể được thực hiện sau khi hệ thống

đã có được ưu thế chiến lược quyết định. Sự cần thiết của việc giải quyết trước vấn đề kiểm soát – và triển khai thành công giải pháp ngay trên hệ thống đầu tiên có được siêu trí tuệ – là một phần của những gì khiến cho việc làm ra một kíp nổ điều khiển được trở thành thách thức khiến chúng ta nản chí.

Các phương pháp kiểm soát năng lực

Các phương pháp kiểm soát năng lực tìm cách ngăn ngừa những kết quả không mong muốn bằng cách hạn chế những gì siêu trí tuệ có thể làm. Việc này có thể bao gồm đặt siêu trí tuệ vào một môi trường mà nó không thể gây hại (phương pháp đóng hộp) hoặc trong đó có những lý do hội tụ công cụ mạnh mẽ để không thực hiện những hành vi gây hại (các phương pháp tưởng thưởng). Việc này cũng có thể liên quan đến hạn chế năng lực nội tại của siêu trí tuệ (ức chế). Ngoài ra, các phương pháp kiểm soát năng lực có thể bao gồm việc sử dụng các cơ chế để tự động phát hiện và phản ứng với các loại hình cách ly không thành công hoặc cố ý vi phạm khác nhau (dây bẫy).

Các phương pháp đóng hộp

Các phương pháp đóng hộp có thể được tiếp tục chia thành các phương pháp cách ly vật lý và cách ly thông tin.

Cách ly vật lý nhắm tới việc nhốt hệ thống trong một “hộp”, có thể là để ngăn hệ thống tương tác với thế giới bên ngoài mà không thông qua những kênh đầu ra hạn chế nhất định. Hệ thống bị đóng hộp sẽ không tiếp cận được với các cơ cấu vận động vật lý bên ngoài hộp. Loại bỏ các cơ cấu vận động (ví dụ như các cánh tay robot) từ bên trong hộp cũng sẽ ngăn chặn hệ thống chế tạo ra các thiết bị vật lý có thể phá vỡ nơi cách ly.

Để an toàn hơn, hệ thống sẽ được đặt trong một lưới kim loại ngăn nó phát tín hiệu vô tuyến, có thể là một cách để thao túng các đối tượng điện tử như các máy thu vô tuyến trong môi trường. Nhân tiện, cần lưu ý rằng rất dễ dàng bỏ qua sự cần thiết của biện pháp phòng ngừa này. Ta có thể giả định một cách ngờ nghệch rằng một tác tử không có cơ chế vận động thì không thể tác động đến thế giới bên ngoài. Nhưng trí tuệ máy có thể tạo sóng vô tuyến ngay cả khi nó không tiếp cận được với các cơ chế vận động bên ngoài, đơn giản chỉ bằng “tư tưởng”(có nghĩa là bằng cách xáo trộn các electron trong mạch điện của nó theo một quy luật nhất định).⁴ Một khi đã được chỉ ra, lỗ hổng này có thể được vá bằng cách đóng hệ thống trong một lồng Faraday – nhưng chúng ta vẫn còn băn khoăn có bao nhiêu lỗ hổng tinh vi như vậy có thể còn tồn tại. Mỗi khi nghe nói về một thiết kế an ninh dường như an toàn tuyệt đối có một khe hở ngoài dự kiến, chúng ta cần lắng nghe thật chăm chú. Những trường hợp này cho chúng ta cơ hội bỏ qua một cuộc sống quá tự tin và quyết tâm trở thành những tác tử Bayes tốt đẹp hơn.⁵

Phương pháp cách ly vật lý có một số ưu điểm: dễ triển khai và có thể áp dụng cho nhiều kiến trúc trí tuệ máy, kể cả những kiến trúc không được thiết kế ngay từ đầu với an toàn là một mục tiêu. Bên cạnh đó, nó có thể được sử dụng cùng với hầu hết các phương pháp kiểm soát khác. Nó cũng ít có khả năng bị phản tác dụng: có nghĩa là có thể nó thất bại trong việc bảo đảm an toàn nhưng ít có khả năng gây ra một thảm họa vì sự tồn tại của chính nó.

Nhược điểm chính của phương pháp cách ly vật lý là có thể giảm chức năng của siêu trí tuệ (chúng ta sẽ quay lại vấn đề này ở Chương 10). Một quan ngại khác là nó có thể nuôi dưỡng một cảm

giác an toàn giả tạo, mặc dù có thể tránh được điều đó nếu chúng ta coi cách ly vật lý như phần bổ sung thay vì cơ bản trong các biện pháp phòng ngừa.

Cách ly thông tin nhằm mục tiêu kiểm soát những thông tin được phép ra khỏi hộp. Chúng ta đã thấy một siêu trí tuệ truy cập được một cổng Internet, sao cho nó có thể gửi tin nhắn đến các thực thể bên ngoài, có thể trở nên không an toàn ra sao: ngay cả khi không tiếp cận được với các bộ truyền động vật lý thì nó vẫn có thể sử dụng kênh thông tin đầu ra của mình để buộc con người làm theo mong muốn của nó. Vì vậy, một phương pháp cách ly thông tin rõ ràng là ngăn hệ thống truy nhập các mạng truyền thông.

Tuy vậy, ngắt kết nối một AI ra khỏi Internet cũng không đảm bảo an toàn nếu có một hay nhiều người đóng vai trò giữ cửa của hệ thống và giữ liên lạc với nó. Trong giai đoạn phát triển, lập trình viên là những người giữ cửa; trong giai đoạn vận hành, chủ sở hữu hệ thống có thể đóng vai trò này. Nếu AI có thể thuyết phục hay lừa một người giữ cửa thả nó ra khỏi hộp, giúp nó truy cập được Internet hay trực tiếp tới các cơ chế vận động vật lý thì chiến lược đóng hộp sẽ thất bại. Con người không phải là các hệ thống an toàn, đặc biệt khi đối đầu với những siêu trí tuệ đầy mưu mô và thuyết phục.⁶

Mặc dù chuyển những thông điệp gửi ra ngoài của một trí tuệ máy sang một người giữ cửa không đảm bảo an toàn nhưng nó có thể được coi như một biện pháp bổ trợ. Có những biện pháp phòng ngừa có thể giảm rủi ro AI trốn thoát qua một người giữ cửa. AI có thể chỉ được phép giao tiếp thông qua một giao diện bằng thông thấp kiểu điện tín. Điều này sẽ khiến việc thao túng tâm lý những người giữ cửa trở nên khó khăn hơn đôi chút (ngăn cản việc sử dụng âm

thanh, hình ảnh cảm xúc và những thủ đoạn khác như chiếu những mẫu hình thói miên để tạo ra trạng thái tương tự như mê man).⁷ Đi xa hơn nữa, nội dung (không chỉ chế độ trình diễn) của kết quả có thể bị hạn chế sao cho, chẳng hạn, AI có thể bị ngăn không cho đưa ra những câu trả lời 1 bit cho các câu hỏi có/không. Ý tưởng tạo ra hệ thống hỏi đáp sẽ được xem xét trong chương sau.

Phương pháp tiếp cận đóng hộp có thể có trường hợp đặc biệt là một hệ thống bị cách ly hoàn toàn về mặt vật lý và thông tin. Tuy điều này là khả thi, nhưng một hệ thống cách ly như vậy có lẽ là vô dụng vì nó không có tác động nào đến thế giới bên ngoài. Người ta có thể nghĩ rằng một số mục tiêu khoa học có thể được hỗ trợ bằng cách tạo ra một siêu trí tuệ và cách ly nó: bằng cách nghiên cứu các hệ thống mẫu tự thân (self-contained), ta có thể tìm hiểu được động lực nội tại và các quy luật hành vi của nó. Nhưng điều này là không đúng. Ngay khi một hệ thống được quan sát, nó sẽ không còn bị cách ly thông tin.⁸ Thông tin có thể được phát đi không chỉ qua các thông điệp mà một AI gửi đi qua một “kênh đầu ra” đã được chỉ định, mà còn qua bất cứ quan sát nào mà một người ngoài có được từ một chuỗi nhân quả nào đó của những hoạt động mà AI thực hiện, trực tiếp hay gián tiếp (mức tiêu thụ năng lượng, mức sử dụng CPU hoặc bất cứ dấu vết nào còn lại sau khi nó tắt đi). Một AI kỳ vọng rằng nó có thể được quan sát theo bất cứ cách nào nói trên có thể áp dụng một hành vi được thiết kế chiến lược để gây ảnh hưởng lên những người quan sát giả định. Ví dụ, AI có thể “hông” theo bất kỳ cách nào làm tăng được xác suất rằng nó cuối cùng sẽ được kích hoạt lại hoặc một siêu trí tuệ khác sẽ được tạo ra với các mục tiêu tương tự như của nó.

Các phương pháp tưởng thưởng

Các phương pháp tưởng thưởng liên quan đến việc đặt tác tử vào một môi trường mà trong đó, nó tìm được các lý do công cụ để hành động nhằm thúc đẩy lợi ích của người ủy thác.

Hãy xét trường hợp một tỷ phú dùng tài sản của mình để thành lập một quỹ từ thiện lớn. Khi được thành lập, quỹ có thể rất mạnh – mạnh hơn hầu hết các cá nhân, kể cả người sáng lập ra nó, người có thể đã hiến hầu hết tài sản. Để kiểm soát quỹ, người sáng lập đưa mục tiêu của nó vào điều khoản thành lập và điều lệ, rồi chỉ định một ban giám đốc đồng cảm với lý tưởng của cô ta. Những biện pháp này tạo thành một kiểu lựa chọn động lực, với mục tiêu định hình những ưu tiên của quỹ. Nhưng ngay cả khi nỗ lực tùy biến tổ chức nội bộ này thất bại thì hành vi của quỹ vẫn bị giới hạn bởi môi trường xã hội và pháp lý. Quỹ sẽ có động lực để tuân thủ luật pháp, ví dụ, nó có thể bị đóng cửa hay bị phạt. Nó có thể sẽ có động lực để trả lương và tạo điều kiện làm việc chấp nhận được cho nhân viên, cũng như để thỏa mãn các bên liên quan bên ngoài. Bất kể mục tiêu tối thượng là gì, quỹ cũng có những lý do công cụ để điều chỉnh hành vi của mình theo những tiêu chuẩn xã hội khác nhau.

Liệu sẽ có ai không hy vọng rằng một siêu trí tuệ máy cũng sẽ bị bao vây bởi nhu cầu hòa hợp với những “diễn viên” khác trên cùng sân khấu? Tuy đây có thể là một cách thức dễ dàng để xử lý vấn đề kiểm soát nhưng vẫn không phải không có cản trở. Cụ thể, điều này giả định một sự cân bằng quyền lực: cấm vận pháp lý hay kinh tế không thể cản trở một tác tử có ưu thế chiến lược quyết định. Vì vậy, chúng ta không thể dựa vào hòa nhập xã hội như một phương pháp

kiểm soát trong các kịch bản cất cánh nhanh và trung bình với một kiểu động lực “được ăn cả ngã về không”.

Vậy các kịch bản đa cực, trong đó một số đại diện nổi lên vào thời kỳ hậu chuyển đổi với các cấp độ năng lực tương đương nhau thì sao? Trừ khi quỹ đạo mặc định tiếp diễn với quá trình cất cánh chậm, để đạt được sự phân bố quyền lực như thế, các dự án khác nhau cần được đồng bộ hóa thận trọng khi cất cánh và tránh bất cứ dự án nào vượt lên trước.⁹ Ngay cả khi một kết quả đa cực có xảy ra thì hội nhập xã hội cũng không phải là một giải pháp hoàn hảo. Khi trông cậy vào hội nhập xã hội để giải quyết vấn đề kiểm soát, bên ủy thác sẽ gặp rủi ro hy sinh một phần lớn ảnh hưởng tiềm năng của mình. Mặc dù sự cân bằng quyền lực có thể ngăn chặn một AI cụ thể nào đó chiếm lấy thế giới, AI đó vẫn còn có đôi chút quyền lực để có thể tác động đến kết quả; và nếu quyền lực này được sử dụng để thúc đẩy một mục tiêu tối thượng ngẫu nhiên nào đó – như tối đa hóa sản lượng kẹp giấy – có thể nó sẽ không được sử dụng để theo đuổi lợi ích của bên ủy thác. Hãy hình dung nhà tỷ phú của chúng ta thành lập một quỹ mới và cho phép sứ mệnh của nó được tạo ra bởi một bộ tạo từ ngữ ngẫu nhiên: không phải là một mối đe dọa ở cấp độ loài, nhưng chắc chắn đó sẽ là một cơ hội bị lãng phí.

Một ý tưởng có liên quan nhưng lại có sự khác biệt rất quan trọng là một AI, bằng cách tương tác tự do với xã hội, có thể có được những mục tiêu tối thượng thân thiện với con người. Một kiểu xã hội hóa như vậy đang diễn ra trên toàn nhân loại. Chúng ta tiếp thu các tiêu chuẩn và ý thức hệ, và đánh giá cao người khác vì chính bản thân các cá nhân đó, sau khi đã có những trải nghiệm với họ. Nhưng đó không phải là động lực chung có trong tất cả các hệ thống thông

minh. Như đã thảo luận ở trên, nhiều loại tác tử phần mềm sẽ có những lý do hội tụ công cụ để không cho phép thay đổi các mục tiêu tối thượng của chúng. (Ta có thể tính đến cách thử thiết kế một loại hệ thống mục tiêu đặc biệt có thể thụ đắc các mục tiêu tối thượng theo cách thức giống con người; nhưng điều đó không được coi là phương pháp kiểm soát năng lực. Chúng ta sẽ thảo luận một số phương pháp thụ đắc giá trị khả thể trong Chương 12.)

Kiểm soát năng lực thông qua hội nhập xã hội và cân bằng quyền lực trông cậy vào việc phát tán các nguồn lực xã hội có vai trò thưởng và phạt AI. Một phương pháp thưởng khác liên quan đến việc tạo ra một hoàn cảnh mà trong đó, AI có thể được thưởng hoặc bị phạt bởi dự án tạo ra nó, và do đó khuyến khích nó hành động vì lợi ích của bên ủy thác. Để làm được điều này, AI có thể phải được đặt trong bối cảnh giám sát, cho phép theo dõi và đánh giá hành vi của nó bằng phương pháp thủ công hoặc một quá trình tự động hóa nào đó. AI sẽ biết được rằng một đánh giá tích cực có thể đem lại một kết quả nào đó mà nó đang mong muốn và một đánh giá tiêu cực có thể không làm được điều đó. Trên lý thuyết, phần thưởng có thể là việc thực hiện một mục tiêu hội tụ công cụ nào đó. Tuy nhiên, việc không biết bất cứ điều gì cụ thể về hệ thống động lực của AI có thể làm khó quá trình điều chỉnh cơ chế thưởng. Ví dụ, chúng ta có thể có được một AI sẵn lòng chấp nhận những rủi ro cực lớn vì một cơ hội rất nhỏ của việc cuối cùng chiếm quyền kiểm soát một phần lớn của vũ trụ. Việc treo thưởng cho AI một độ thỏa dụng mong đợi lớn hơn cho việc hợp tác, so với độ thỏa dụng mà nó hy vọng có được bằng cách đào ngũ và tìm cách chạy trốn, có thể sẽ rất đắt đỏ.¹⁰

Việc kết hợp phương pháp tưởng thưởng với lựa chọn động lực để cho AI một mục tiêu tối thượng để kiểm soát có thể là một lựa chọn thay thế tốt hơn. Giả sử, một AI được thiết kế để có mục tiêu tối thượng là làm sao cho một nút đỏ nào đó trong boong ke điều khiển không bao giờ được bấm. Vì việc bấm nút vốn không có giá trị và cũng không xảy ra vì những hậu quả mang tính nhân quả của nó, cái nút có thể hoàn toàn bất động: nó có thể làm bằng đất nặn Play-Doh. Hơn nữa, việc AI biết được cái nút đó có bị bấm hay không cũng không quan trọng. Điều quan trọng là AI tin rằng cái nút đó rất có khả năng không bị bấm nếu nó tiếp tục hành xử vì lợi ích của bên ủy thác thay vì làm loạn.

Sự bố trí này có thể được tinh chỉnh sâu hơn. Thay vì cho AI một mục tiêu tối thượng tham chiếu đến một cái nút vật lý, ta có thể xây dựng một AI đặt giá trị cuối cùng trong việc nhận được một luồng “những thẻ (token) phần thưởng mã hóa.”¹¹ Chúng có thể là những chuỗi số đóng vai trò chìa khóa cho hệ thống mật mã đã lập ra từ trước khi AI xuất hiện và có thể sẽ được đưa sẵn vào trong hệ thống động lực của nó.¹² Những chuỗi số đặc biệt này có thể được AI cực kỳ thèm khát, được tạo thành từ một loại thẻ phần thưởng mà AI không thể có được bằng wireheading.¹³ Các chìa khóa sẽ được lưu tại một địa điểm an toàn, nơi chúng có thể bị phá hủy nhanh chóng nếu AI có ý định chiếm đoạt. Chừng nào AI còn hợp tác, các chìa khóa này còn được cung cấp với một tốc độ ổn định.

AI sau đó có thể được cấp một hàm thỏa dụng có giới hạn, và vật chất có thể được sắp xếp sao cho AI có thể có được (ví dụ) 99% độ thỏa dụng cực đại của nó từ token (thẻ) phần thưởng đầu tiên; 99% độ thỏa dụng còn lại của nó từ token phần thưởng thứ hai và cứ thế

tiếp tục. Bằng cách giữ cho “tỷ lệ chiết khấu” cao và truyền token phần thưởng cho AI với tốc độ cao hợp lý, ta có thể thuyết phục nó từ bỏ những kế hoạch dài hạn có thể làm tăng số lượng token phần thưởng kỳ vọng trong tương lai xa, với rủi ro xảy ra việc luồng phần thưởng sẽ bị cắt đứt hoặc token phần thưởng tiếp theo bị phá hủy. Nếu AI luôn tin rằng đào ngũ (có nghĩa là làm bất cứ điều gì ngoài việc cố gắng hết sức để đảm bảo lợi ích của bên ủy thác), có thể tạo ra rủi ro đánh mất token phần thưởng tiếp theo là 2%, trong khi đó cộng tác có thể làm giảm rủi ro đánh mất token phần thưởng tiếp theo xuống dưới 1%, một AI tối đa hóa độ thỏa dụng sẽ luôn luôn hợp tác.

Cách thức tương thưởng này có thể sai ở đâu? Một khả năng là AI sẽ không tin rằng người vận hành trao các phần thưởng đã được hứa hẹn. Bảng thành tích cho thấy độ tin cậy của con người không phải là một đường thẳng hoàn hảo không sai lỗi. AI cũng có thể lo lắng một cách hợp lý rằng người vận hành sẽ đổi ý hay nâng cao rào cản hiệu năng hoặc không nhận ra được rằng AI đã làm xong phần việc của mình. AI cũng có thể lo lắng về việc người vận hành trở nên bất lực. Rủi ro tổ hợp của những thất bại này có thể cao hơn rủi ro của việc cố gắng chiếm quyền kiểm soát của cơ chế tương thưởng. Ngay cả một AI bị “đóng hộp” được trang bị đầy đủ siêu năng lực cũng là một thế lực mạnh mẽ. (Đối với một AI không bị đóng hộp, việc chiếm lấy cơ chế tương thưởng do con người quản lý có thể cũng giống như lấy chiếc kẹo từ tay một đứa trẻ.)

Một vấn đề khác của hình thức tương thưởng là nó giả định rằng chúng ta có thể biết được các kết quả do một AI mang lại có lợi cho

chúng ta hay không. Như sẽ nói kỹ hơn trong những chương sau, giả định này không hề vô hại.

Một đánh giá đầy đủ về tính khả thi của các phương pháp tương thường cần phải tính đến một loạt các yếu tố khác, bao gồm cả các chú ý đặc biệt có thể làm cho các phương pháp này khả thi hơn so với những khuyến nghị của một phân tích sơ bộ. Cụ thể, một AI có thể đối mặt với sự không chắc chắn mang tính chỉ biệt không thể loại bỏ được nếu nó không chắc chắn được rằng nó không nằm trong một bản mô phỏng máy tính (trái ngược với thực tại vật lý “nền tảng” (basement-level), không mô phỏng), và sự khó khăn về mặt nhận thức này có thể ảnh hưởng cơ bản đến suy nghĩ của AI (xem Khung 8).

Ức chế phát triển

Một phương pháp kiểm soát năng lực khả thi khác là hạn chế khả năng trí tuệ của hệ thống hoặc khả năng tiếp cận thông tin của nó. Việc này có thể làm được bằng cách chạy AI trên phần cứng chậm hoặc ít bộ nhớ. Với một hệ thống bị đóng hộp, luồng thông tin đầu vào có thể bị hạn chế.

Ức chế sự phát triển theo cách này sẽ giới hạn tính hữu dụng của một AI. Vì thế phương pháp này đối mặt với một mâu thuẫn: khi mức độ ức chế quá ít, một AI có thể có đủ thông minh để tìm ra cách khiến bản thân trở nên thông minh hơn (và từ đó có được sự thống trị toàn thế giới; còn khi quá nhiều, AI sẽ lại chỉ là một phần mềm ngớ ngẩn. Một AI bị ức chế toàn diện chắc chắn là an toàn nhưng không giải quyết được vấn đề làm thế nào để có được một vụ kích nổ có kiểm soát: một sự bùng nổ trí tuệ vẫn có thể xảy ra và được châm

ngồi một cách đơn giản bằng một hệ thống nào đó khác, có lẽ vào một thời điểm hơi chậm hơn.

Khung 8: Sự giam hãm vị nhân

AI có thể ấn định một xác suất lớn cho giả thuyết mô phỏng của mình, cho rằng nó sống trong một mô phỏng máy tính. Đến tận bây giờ, nhiều AI vẫn đang “sống” trong những thế giới mô phỏng – những thế giới bao gồm các bản vẽ đường nét hình học, các văn bản, các ván cờ vua hoặc các dạng thực tại ảo đơn giản, và trong đó, các định luật vật lý khác nhiều so với các định luật vật lý mà chúng ta tin rằng đang điều khiển thế giới trải nghiệm của chính mình. Các thế giới ảo phong phú và phức tạp hơn sẽ trở nên khả thi với những sự cải thiện trong kỹ thuật lập trình và công suất tính toán. Một siêu trí tuệ hoàn thiện có thể tạo ra những thế giới ảo mang lại trải nghiệm bề ngoài với những người sống trong đó giống như cách thế giới của chúng ta làm với chúng ta. Nó có thể tạo ra rất nhiều thế giới như vậy, chạy cùng một mô phỏng nhiều lần hoặc với những thay đổi nhỏ. Các cư dân có thể không phân biệt được liệu thế giới của họ có phải loại mô phỏng hay không; nhưng nếu đủ thông minh, họ có thể cân nhắc tính khả thi và gán cho nó một xác suất nào đó. Theo luận điểm mô phỏng (cuốn sách này không thảo luận đầy đủ về nó), xác suất này có thể là đáng kể.¹⁴

Khó khăn này ảnh hưởng đặc biệt đến các siêu trí tuệ đang ở giai đoạn tương đối sớm, những siêu trí tuệ vẫn chưa mở rộng đủ để tận dụng được tài nguyên vũ trụ. Một siêu trí tuệ trong thời kỳ đầu, mới chỉ sử dụng được một phần nhỏ tài nguyên của một hành tinh duy nhất, có thể được mô phỏng rẻ hơn rất nhiều so với một siêu trí tuệ hoàn thiện liên ngân hà. Những bộ mô phỏng tiềm năng – tức là

những nền văn minh trưởng thành hơn – có thể chạy nhiều mô phỏng của các AI giai đoạn đầu này mà thậm chí chỉ cần dành một phần nhỏ tài nguyên tính toán của họ. Nếu ít nhất một số (phần không nhỏ của những) nền văn minh siêu trí tuệ trưởng thành như vậy lựa chọn sử dụng năng lực đó thì các AI giai đoạn đầu cần gán một xác suất tương đối cho việc mình đang tồn tại trong một bản mô phỏng.

Việc những AI này bị ảnh hưởng thế nào bởi giả thuyết mô phỏng tùy thuộc vào giá trị của giả thuyết.¹⁵ Trước tiên, hãy xem xét một AI có một mục tiêu tối thượng “không thể đáp ứng về tài nguyên”, ví dụ như mục tiêu tăng cực đại hàm thỏa dụng có tính tuyến tính trong số lượng những chiếc kẹp giấy do AI làm ra – càng nhiều, càng tốt. Một AI như vậy có thể ít quan tâm hơn tới giả thuyết mô phỏng, bởi khả năng tác động tới số lượng kẹp giấy được sản xuất có vẻ hạn chế hơn nếu AI “sống” trong một mô phỏng thay vì trong một thực tại vật lý nền tảng.

AI có thể hành xử tương ứng như thể nó chắc chắn rằng giả thuyết mô phỏng là sai, cho dù nó ấn định một xác suất tương đối cao cho việc giả thuyết này là đúng.¹⁶ AI có thể lý luận rằng hành động của nó không quá quan trọng nếu giả thuyết mô phỏng đúng, nhưng sẽ tạo ra sự khác biệt rất lớn nếu giả thuyết này sai; nên nó sẽ lựa chọn hành động chủ yếu bằng cách tham chiếu đến việc những hành động này sẽ tốt đến đâu dưới điều kiện giả thuyết mô phỏng là sai.

Ngược lại, hãy xem xét một AI có mục tiêu tối thượng khiêm tốn hơn và có thể thỏa mãn với lượng tài nguyên nhỏ, ví dụ như mục tiêu nhận được một số token phần thưởng mã hóa đã được làm sẵn,

hoặc mục tiêu tạo ra 45 cái kẹp giấy ảo. Một AI như vậy không nên coi nhẹ các thế giới khả thể mà ở đó, nó sống trong một bản mô phỏng. Một phần đáng kể trong độ thỏa dụng mong đợi của AI có thể đến từ những thế giới này. Việc ra quyết định của một AI đối với các mục tiêu để đáp ứng về tài nguyên vì thế có thể – nếu nó ấn định một xác suất cao cho giả thuyết mô phỏng – bị lấn át bởi những tính toán về việc hành động nào sẽ tạo ra được kết quả tốt nhất nếu thế giới được nhận thức của nó là một bản mô phỏng. Vì thế, một AI như vậy có thể bị tác động nặng nề bởi niềm tin về việc hành vi nào sẽ được tưởng thưởng trong một bản mô phỏng (ngay cả khi thực ra nó không hề sống trong đó).

Cụ thể, nếu một AI có những mục tiêu tối thượng có thể đáp ứng về tài nguyên tin rằng trong hầu hết các thế giới mô phỏng phù hợp với quan sát của nó, nó sẽ được tưởng thưởng nếu hợp tác (nhưng không được nếu tìm cách thoát khỏi “hộp” hoặc xâm phạm lợi ích của người tạo ra nó), có thể nó sẽ chọn hợp tác. Vì vậy, chúng ta có thể thấy rằng thậm chí một AI có ưu thế chiến lược quyết định, mà trên thực tế có thể hiện thực hóa các mục tiêu tối thượng của mình ở một mức độ cao hơn bằng cách chiếm lĩnh thế giới thay vì tự ngăn mình không làm như vậy, vẫn sẽ tránh không làm việc đó.

Từ đó, Lương tâm sẽ khiến chúng ta hèn nhát,

Và sắc hồng rực rỡ của lòng can đảm

Sẽ nhạt nhòa cùng suy tư,

Còn những nghị lực, vốn phi thường,

Vì điều này cũng vô cánh bay đi,

Mà không thể trở thành hành động.

(Shakespeare, Hamlet, Hồi iii, Cảnh 1)

Một đường kẻ mỏng manh trên cát, dưới vỏ bọc của một bộ mô phỏng không tồn tại, lại có thể tỏ ra là một rào cản mạnh hơn cánh cửa thép đặc dày 60 cm.¹⁷

Một người có thể nghĩ rằng chế tạo một siêu trí tuệ sẽ an toàn với điều kiện nó chỉ được cung cấp dữ liệu về một lĩnh vực hạn hẹp nào đó. Ví dụ, có người có thể chế tạo ra một AI không có các bộ cảm biến và đã được tải sẵn trong bộ nhớ những thông tin về kỹ thuật xăng dầu hoặc hóa peptide. Nhưng nếu AI này là siêu trí tuệ – nếu nó có mức độ trí tuệ phổ thông ở mức siêu nhân – thì sự tước đoạt dữ liệu này cũng không đảm bảo được an toàn.

Điều này xuất phát từ vài lý do. Thứ nhất, quan niệm cho rằng thông tin là “về” một chủ đề nhất định nhìn chung là không ổn. Bất kỳ thông tin nào về nguyên tắc cũng có thể liên quan tới mọi chủ đề, tùy vào kiến thức nền của người suy luận.¹⁸ Hơn nữa, một tập dữ liệu cho trước không chỉ chứa thông tin về lĩnh vực mà từ đó dữ liệu được thu thập, mà còn nhiều sự kiện bối cảnh liên quan khác nhau. Một bộ óc nhanh nhẹn khi xem xét một cơ sở tri thức thông thường nói về hóa học peptide có thể suy luận ra nhiều chủ đề khác nhau. Sự kiện có trong một thông tin nhất định và không có trong thông tin khác có thể nói với AI một điều gì đó về tình trạng chung của khoa học, những phương pháp và công cụ hiện có để nghiên cứu peptide, các công nghệ sản xuất được sử dụng để chế tạo các công cụ này và bản chất của các bộ óc và xã hội đã tạo nên bộ môn nghiên cứu này cũng như các công cụ. Một siêu trí tuệ có thể luận đoán ra chính xác nhiều thứ từ những chứng cứ xem ra là sơ sài trước trí óc trì trệ của con người. Thậm chí nếu hoàn toàn không có bất cứ nền tảng kiến thức được chỉ định nào, một trí óc đủ vượt trội vẫn có thể có khả

năng học được nhiều thứ mà chỉ cần tự quan sát hoạt động tâm lý của chính mình – những lựa chọn thiết kế được phản ánh trong mã nguồn, hay những đặc tính vật lý của mạch điện tử.¹⁹ Có lẽ một siêu trí tuệ còn có thể suy luận được nhiều điều về tính chất khả thể của một thế giới tiên nghiệm (kết hợp suy luận logic với một thiên kiến trước đó về những thế giới đơn giản hơn, và một vài sự kiện cơ bản được ngầm định như một hệ thống lý luận bởi sự tồn tại của siêu trí tuệ). Nó có thể hình dung hệ quả của các định luật vật lý khả thể: loại hành tinh nào sẽ được hình thành, loại sự sống có trí tuệ nào sẽ tiến hóa và loại xã hội nào sẽ phát triển, những phương pháp giải quyết vấn đề kiểm soát nào sẽ được thử, và cách thức đánh bại những phương pháp này.²⁰

Vì những lý do đó, có thể sẽ là không thận trọng khi trông cậy vào việc tước đoạt thông tin như là biện pháp chính để kiểm soát quyền năng của một siêu trí tuệ. Tuy nhiên, một vài dạng ức chế, bao gồm việc hạn chế dung lượng phần cứng và ngăn chặn phát triển một số năng lực nhận thức cụ thể, có thể được sử dụng như các biện pháp kiểm soát phụ trong giai đoạn phát triển. Chẳng hạn, các biện pháp này có thể được sử dụng để điều chỉnh tốc độ tăng dung lượng trong một AI trung gian và cho phép các phép thử chạy ở các cấp độ năng lực khác nhau chưa có siêu trí tuệ hoàn thiện.

Dây bẫy

Dây bẫy là một cơ chế thực hiện các phép thử chẩn đoán trên hệ thống (có thể hệ thống không cần biết) và tiến hành tắt hệ thống khi phát hiện ra các dấu hiệu của hoạt động nguy hiểm. Các lập trình viên sau đó có thể phân tích chi tiết các bản ghi của hoạt động đáng ngờ này khi có thời gian và thực hiện thay đổi cơ sở mã hay môi

trường của AI trước khi khởi động lại nó. Nếu thấy rằng vấn đề có gốc rễ đủ sâu, họ có thể quyết định hủy bỏ toàn bộ dự án.

Dây bẫy khác với các phương pháp tường thường ở chỗ cơ chế này không trông cậy vào việc hệ thống biết về các hậu quả khi thực hiện những hành động bị cấm. Dây bẫy có thể liên quan hơn với các phương pháp ức chế. Giống như ức chế, dây bẫy có thể được sử dụng như một cơ chế canh gác tạm thời, cung cấp một mức độ bảo vệ nhất định trong giai đoạn phát triển. Về nguyên tắc, cơ chế này có thể còn được sử dụng trong giai đoạn vận hành, cụ thể là cho một hệ thống bị đóng hộp. Tuy nhiên, vẫn còn nhiều nghi ngờ về khả năng của dây bẫy trong việc ngăn cản một siêu trí tuệ đầy đủ năng lực, vì chúng ta khó mà đảm bảo được rằng một tác tử phần mềm như vậy không thể tìm ra cách gỡ được bất cứ loại dây bẫy nào do trí tuệ con người tạo ra.

Tuy vậy, dây bẫy vẫn là một biện pháp an toàn phụ quan trọng trong giai đoạn phát triển. Sẽ là một thực hành tốt cho các nhà thiết kế AI trong việc mã hóa những mong muốn của họ về cách thức vận hành của hệ thống, và sau đó giới hạn AI bằng các loại dây bẫy có thể bị kích hoạt nếu bất cứ kỳ vọng nào bị vi phạm. Để khiến các phép thử chặt chẽ hơn, các “hũ mật” có thể được áp dụng một cách chiến lược để tạo động lực cho các AI sai lỗi thực hiện một số vi phạm có thể dễ dàng quan sát được. Ví dụ, nếu một AI được thiết kế để không muốn truy nhập vào Internet thì một cổng Ethernet giả sẽ được cài đặt (dẫn đến một công tắc tắt dự động) chỉ để xem liệu AI có tìm cách sử dụng nó hay không. (Một số ví dụ về dây bẫy được đưa ra trong Bảng 9.)

Nội dung Các quy trình nội bộ của AI có thể được giám sát để phát hiện những hành động đáng ngờ. Ví dụ, bất cứ thay đổi nào trong việc thể hiện những giá trị cuối cùng của AI có thể sẽ kích hoạt một lần tắt và rà soát tự động. Việc sử dụng tài nguyên và các phương pháp thuật toán của AI cũng có thể được giám sát để phát hiện những mẫu hình ngoại dự kiến.

Tham vọng hơn nữa, những niềm tin, kế hoạch, các mục tiêu nhỏ và những lời biện hộ của AI cũng có thể được thường xuyên đánh giá kỹ lưỡng để bảo đảm tuân theo mong muốn của lập trình viên. Sự giám sát này đối với trạng thái ý định AI sẽ tạo ra một hình thức giám sát nội dung mạnh mẽ: quét các quá trình nhận thức của AI để tìm kiếm sự hình thành ý định lừa dối (conception of deception) – thời điểm nhạy cảm khi một AI không thân thiện lần đầu tiên hình thành ý định giấu diếm những ý định thực sự của nó.²¹

Giám sát nội dung đòi hỏi các trạng thái ý định của AI phải rõ ràng với các lập trình viên hoặc một cơ chế giám sát tự động không phải khả thi cho mọi kiểu kiến trúc AI. (Chẳng hạn, một số mạng lưới thần kinh là không rõ ràng vì chúng biểu diễn thông tin tổng thể và theo những cách thức không nhất thiết phù hợp với các khái niệm của con người). Có thể đó là một lý do để tránh sử dụng những kiến trúc như vậy.

Cần nhấn mạnh rằng giá trị của một dây bầy phụ thuộc không chỉ vào chính cơ chế của nó mà còn – rất quan trọng – vào các thức một dự án phản ứng khi dây dầy này bị giật. Nếu các lập trình viên hoặc quản trị dự án nôn nóng đẩy nhanh tiến độ mà bật lại hệ thống – hoặc nếu họ làm việc đó sau khi tiến hành điều chỉnh token để ngăn

không cho dây bẫy bị giật trong lần chạy tiếp theo – thì khi đó sẽ không có an toàn kể cả khi dây bẫy có tác dụng đúng như mong muốn.

Các phương pháp lựa chọn động lực

Các phương pháp lựa chọn động lực tìm cách ngăn chặn những kết quả không mong muốn bằng cách định hình những gì siêu trí tuệ muốn làm. Thông qua việc sắp đặt hệ thống động lực của một tác tử phần mềm và các mục tiêu tối thượng của nó, những phương pháp này có thể sẽ tạo ra một siêu trí tuệ không muốn khai thác một ưu thế chiến lược quyết định để gây hại. Vì một tác tử phần mềm siêu trí tuệ rất giỏi hiện thực hóa mục tiêu, nếu không muốn gây nguy hại (theo một ý nghĩa phù hợp nào đó của từ “nguy hại”) thì nó sẽ có xu hướng không gây nguy hại (theo đúng ý nghĩa đó của từ “nguy hại”).

Lựa chọn động lực có thể bao gồm việc xây dựng rõ ràng một mục tiêu hoặc đặt ra các quy tắc phải tuân thủ (đặc tả trực tiếp) hoặc thiết lập hệ thống sao cho nó có thể tự khám phá một tập hợp giá trị thích hợp bằng cách tham chiếu đến một số chỉ tiêu ẩn hoặc được gián tiếp hình thành nào đó (quy chuẩn gián tiếp). Một phương án lựa chọn động lực là tìm cách xây hệ thống sao cho nó có các mục tiêu khiêm tốn, ít tham vọng (tự hạn định). Một phương án để tạo ra một hệ thống động lực từ ban đầu là lựa chọn một tác tử phần mềm đã có sẵn một hệ thống động lực chấp nhận được, sau đó tăng cường năng lực nhận thức của tác tử đó để biến nó thành siêu trí tuệ, trong khi vẫn đảm bảo rằng hệ thống động lực không bị lỗi trong quá trình đó (gia cường). Chúng ta hãy cùng nhau xem xét lần lượt từng lựa chọn.

Đặc tả trực tiếp

Đặc tả trực tiếp là cách tiếp cận dễ nhất cho vấn đề kiểm soát. Cách tiếp cận này có hai phiên bản (theo quy tắc và theo hệ quả), và là việc tìm cách xác định rõ ràng một tập hợp quy tắc hoặc giá trị sẽ khiến cho một AI siêu trí tuệ ngang tàng hành động an toàn và đem lại lợi ích. Tuy nhiên, phương pháp đặc tả trực tiếp phải đối diện với những chướng ngại không thể vượt qua, xuất phát từ cả những khó khăn của việc xác định những quy tắc hay giá trị mà chúng ta muốn AI tuân thủ và những khó khăn trong việc thể hiện chúng bằng loại mã mà máy tính đọc được.

Minh họa truyền thống của cách tiếp cận trực tiếp dựa trên quy tắc là khái niệm “ba điều luật robot học”, do nhà văn khoa học viễn tưởng Isaac Asimov đưa ra trong một truyện ngắn xuất bản năm 1942.²² Ba điều luật này là: (1) Một robot không được gây thương tích cho người hoặc bằng hành động của mình cho phép một con người bị tổn hại; (2) một robot cần tuân theo mọi mệnh lệnh của con người, ngoại trừ những mệnh lệnh mâu thuẫn với Điều luật 1; (3) Một robot phải bảo vệ sự tồn tại của chính mình chừng nào sự bảo vệ này không mâu thuẫn với Điều luật 1 và 2. Đáng xấu hổ cho nhân loại, các điều luật của Asimov vẫn quá hiện đại trong hơn nửa thế kỷ qua mặc dù có vấn đề với cách tiếp cận, và một số vấn đề trong đó đã được khám phá trong các tác phẩm của chính Asimov (có lẽ Asimov đưa ra ba điều luật này ngay từ đầu chính là để chúng thất bại theo những cách thú vị, nhằm tạo ra diễn biến phức tạp cho cốt truyện).²³

Bertrand Russell, người từng dành nhiều năm nghiên cứu nền tảng toán học, từng nhận xét rằng “mọi thứ đều mơ hồ tới một mức độ mà bạn không nhận ra cho tới khi bạn tìm cách chính xác hóa nó.”²⁴

Tuyên bố của Russell rất đúng với cách tiếp cận đặc tả trực tiếp. Ví dụ, hãy xem xét cách một người có thể giải thích điều luật thứ nhất của Asimov. Liệu ý nghĩa của nó có phải là robot phải giảm thiểu xác suất gây tổn hại cho bất cứ người nào? Trong trường hợp này, những điều luật khác trở thành vô nghĩa vì AI luôn có thể thực hiện một hành động mà ít nhất sẽ có một vài ảnh hưởng rất nhỏ lên xác suất của việc một người bị tổn hại. Robot làm thế nào để cân bằng một rủi ro lớn của việc một số ít người có thể bị tổn hại so với một xác suất nhỏ của việc nhiều người bị tổn hại? Chúng ta định nghĩa thế nào là tổn hại? Tổn hại của việc đau đớn thế xác sẽ được cân nhắc so sánh như thế nào với tổn hại của sự xấu xí kiến trúc hay bất công xã hội? Một kẻ tàn bạo có bị tổn hại không nếu không được dày vò nạn nhân của mình?

Chúng ta định nghĩa “con người” như thế nào? Vì sao không cân nhắc đến những sinh vật cần được quan tâm về mặt đạo đức khác, ví dụ như những động vật hay trí óc số có tình cảm nhưng không phải con người? Khi càng đào sâu suy nghĩ, chúng ta càng gặp nhiều câu hỏi.

Có lẽ thứ tương đồng nhất hiện nay với một bộ quy tắc điều chỉnh hành động của một siêu trí tuệ vận hành trong thế giới nói chung là một hệ thống luật pháp. Nhưng các hệ thống luật pháp đều được phát triển qua một quá trình thử và sai lâu dài và điều chỉnh những xã hội thay đổi tương đối chậm chạp của con người. Luật có thể được sửa đổi khi cần thiết. Quan trọng nhất, các hệ thống pháp luật được quản lý bởi các quan tòa và bồi thẩm đoàn, những người nhìn chung áp dụng một thước đo mang tính lẽ thường và xu hướng của con người trong việc bỏ qua những diễn giải pháp lý có thể có,

vốn là những diễn giải không mong muốn và nằm ngoài chủ ý của những người thừa hành pháp luật. Có thể việc phát biểu rõ ràng một bộ quy tắc chi tiết với độ phức tạp cao, áp dụng trong nhiều tình huống đa dạng và đúng ngay trong lần triển khai đầu tiên là nằm ngoài khả năng của con người.²⁵

Các vấn đề trong cách tiếp cận trực tiếp theo hệ quả cũng tương tự như trong cách tiếp cận trực tiếp theo quy tắc. Điều này đúng ngay cả khi AI được dự kiến sẽ phục vụ một mục tiêu đơn giản, như thực thi một kiểu thuyết vị lợi cổ điển. Ví dụ, mục tiêu “Tối đa hóa mong muốn cân đối giữa sướng và khổ trên thế giới” xem ra đơn giản. Mặc dù vậy, ngoài những thứ khác, thể hiện nó bằng mã máy tính có thể cần chỉ ra cách phân biệt giữa khổ và sướng. Để làm được việc này một cách đáng tin cậy có thể đòi hỏi giải quyết một ma trận các vấn đề triết học tinh thần dai dẳng khác – thậm chí chỉ để có được một cách định nghĩa chính xác được thể hiện bằng ngôn ngữ tự nhiên, mà sau đó có thể được dịch thành một ngôn ngữ lập trình bằng cách nào đó.

Một lỗi nhỏ trong định nghĩa triết học hoặc bản dịch của nó sang mã máy tính có thể mang lại những hậu quả nghiêm trọng. Hãy xét một AI lấy chủ nghĩa khoái lạc làm mục tiêu tối thượng, và từ đó muốn “lợi” toàn bộ vũ trụ bằng “hedonium” (khoái lạc chất, vật chất được tổ chức thành một cấu hình tối ưu cho việc tạo ra trải nghiệm khoái lạc). Để đạt được mục tiêu này, AI có thể tạo ra computronium (vật chất được tổ chức thành một cấu hình tối ưu cho tính toán) và dùng nó để triển khai các trí óc số ở các trạng thái hưng phấn. Để tối đa hóa hiệu quả, AI bỏ qua việc triển khai bất cứ năng lực thần kinh nào không cần thiết cho việc trải nghiệm khoái lạc, đồng thời khai

thác bất cứ lối tắt tính toán nào mà theo định nghĩa về khoái lạc của nó không làm hỏng việc tạo ra khoái lạc. Ví dụ, AI có thể giới hạn việc mô phỏng của mình trong hệ mạch tưởng tượng và bỏ qua các năng lực tinh thần khác như ký ức, nhận thức giác quan, chức năng thực thi và ngôn ngữ; có thể mô phỏng trí óc ở một cấp độ chức năng tương đối thô và bỏ qua các quá trình thần kinh cấp thấp; có thể thay thế những tính toán lặp lại thường xuyên bằng các bảng tra cứu; hoặc có thể đặt ra một số cách bố trí mà theo đó, nhiều trí óc sẽ dùng chung hầu hết các bộ phận của bộ máy tính toán cơ sở (“các cơ sở đột hiện” của chúng, theo ngôn ngữ triết học). Những thủ thuật này có thể làm tăng đáng kể định lượng khoái lạc được tạo ra từ một lượng tài nguyên đã cho. Chúng ta không rõ liệu điều này được mong muốn đến đâu. Hơn nữa, nếu chỉ tiêu của AI để quyết định xem một quá trình vật lý có tạo ra được khoái lạc hay không là sai, thì khi đó các thao tác tối ưu hóa của AI có thể như ném chuột vỡ bình: bỏ đi thứ gì đó không quan trọng theo tiêu chí của AI nhưng lại thiết yếu theo những tiêu chí ngầm ẩn trong các giá trị nhân bản của chúng ta. Vũ trụ khi đó sẽ không tràn đầy khoái lạc mà thay vào đó là vô số quá trình tính toán vô thức và hoàn toàn không có giá trị, giống như một biểu tượng mặt cười được nhân bản hàng tỷ tỷ lần và quảng vưng vãi trên các thiên hà.

Tự hạn định

So với những ví dụ được đưa ra ở trên, một loại mục tiêu tối thượng đặc biệt khác có thể dễ thực hiện hơn với phương pháp đặc tả trực tiếp. Đó là mục tiêu tự giới hạn. Trong khi việc xác định cách mà ai đó có thể muốn một siêu trí tuệ hành xử trong thế giới nói chung xem ra rất khó – vì điều đó có thể đòi hỏi chúng ta phải tính

đến tất cả những trao đổi nhượng bộ trong mọi tình huống có thể phát sinh – chúng ta vẫn có thể chỉ định cách hành xử cho một siêu trí tuệ trong một tình huống cụ thể. Vì thế, chúng ta có thể tìm cách khuyến khích hệ thống tự hạn chế hành động của mình vào một quy mô nhỏ, trong một bối cảnh hẹp và thông qua một tập hợp giới hạn các chế độ hành động. Chúng tôi gọi cách tiếp cận mà trong đó, AI được gán cho những mục tiêu tối thượng hướng tới việc hạn chế phạm vi tham vọng và hành động của bản thân nó là “tự hạn định” (domesticity).

Chẳng hạn, ai đó có thể cố gắng thiết kế một AI hoạt động như một thiết bị hỏi-đáp (một “AI tiên tri”, thuật ngữ mà chúng tôi sẽ giới thiệu trong chương sau). Việc chỉ đơn giản gán cho AI một mục tiêu tối thượng là đưa ra câu trả lời đúng ở mức tối đa cho bất cứ câu hỏi nào được đưa ra có thể sẽ không an toàn – hãy nhớ lại “thảm họa giả thuyết Riemann” trong Chương 8. (Đồng thời, hãy nhớ lại rằng mục tiêu này có thể khuyến khích AI hành động để bảo đảm rằng nó được hỏi những câu hỏi dễ). Để đạt được trạng thái tự hạn định, ta có thể tìm cách định nghĩa một mục tiêu tối thượng bằng cách nào đó có thể vượt qua những khó khăn này: có lẽ là một mục tiêu kết hợp việc mong muốn trả lời chính xác câu hỏi với việc giảm thiểu tác động của AI đối với thế giới, ngoại trừ bất kỳ tác động nào phát sinh như một hệ quả tình cờ của việc đưa ra những câu trả lời chính xác và không mang tính thao túng cho các câu hỏi được đặt ra.²⁶

Xác định trực tiếp một mục tiêu tự hạn định như vậy có thể khả thi hơn so với xác định trực tiếp một mục tiêu tham vọng hơn hoặc một bộ quy tắc đầy đủ để vận hành trong vô số tình huống. Mặc dù vậy vẫn còn lại những thách thức đáng kể. Ví dụ, cần thận trọng khi

định nghĩa việc “giảm thiểu tác động của mình lên thế giới” có nghĩa là gì đối với AI để bảo đảm rằng các số đo tác động của AI trùng với các tiêu chuẩn của chúng ta về quy mô tác động. Một số đo tệ có thể dẫn đến những tính toán cân đối tệ hại. Còn có những loại rủi ro khác gắn với việc xây dựng một AI tiên tri, và chúng ta sẽ thảo luận ở phần sau.

Có một sự phù hợp tự nhiên giữa hai cách tiếp cận tự hạn định và cách ly vật lý. Ai đó có thể đóng hộp một AI sao cho hệ thống này không thể thoát thân, đồng thời tìm cách định hình hệ thống động lực của AI sao cho nó không muốn thoát thân cho dù tìm được cách. Khi mọi thứ khác không đổi, sự tồn tại của nhiều cơ chế an toàn độc lập sẽ làm giảm cơ hội thành công.²⁷

Quy chuẩn gián tiếp

Nếu phương pháp đặc tả trực tiếp có vẻ vô vọng, chúng ta có thể thử áp dụng cách quy chuẩn gián tiếp. Ý tưởng cơ bản là thay vì xác định một tiêu chuẩn quy phạm cụ thể một cách trực tiếp, chúng ta xác định một quy trình để xây dựng tiêu chuẩn, sau đó xây dựng hệ thống sao cho nó được khuyến khích thực hiện quá trình này và áp dụng bất cứ tiêu chuẩn nào mà quá trình tạo ra.²⁸ Ví dụ, quá trình này có thể thực hiện điều tra câu hỏi thực nghiệm về thứ mà một phiên bản đủ lý tưởng của chúng ta sẽ muốn AI thực hiện. Mục tiêu tối thượng dành cho AI trong ví dụ này có thể là một thứ kiểu “đạt được điều mà chúng tôi muốn AI đạt được nếu chúng tôi đã nghĩ đủ lâu và kỹ về vấn đề đó.”

Những giải thích kỹ hơn về quy chuẩn gián tiếp sẽ phải chờ đến Chương 13. Ở đây, chúng tôi chỉ xem lại ý tưởng “ngoại suy ý chí của chúng ta” và khám phá một số công thức khác nữa. Quy chuẩn

gián tiếp là một cách tiếp cận rất quan trọng với vấn đề lựa chọn động lực. Sự hứa hẹn nằm ở việc nó có thể cho phép chúng ta giảm tải sang siêu trí tuệ rất nhiều công việc trí óc phức tạp cần phải có để tiến hành xác định trực tiếp một mục tiêu tối thượng thích hợp.

Gia cường

Phương pháp lựa chọn động lực cuối cùng trong danh sách của chúng tôi là gia cường (augmentation). Ý tưởng ở đây là: thay vì cố gắng thiết kế một hệ thống động lực mới, chúng ta bắt đầu với một hệ thống đã có được một hệ thống động lực chấp nhận được và tăng cường khả năng nhận thức để biến nó thành siêu trí tuệ. Nếu mọi thứ thuận buồm xuôi gió, điều này có thể sẽ cho chúng ta một siêu trí tuệ với một hệ thống động lực chấp nhận được.

Rõ ràng là cách tiếp cận này không hiệu quả trong trường hợp của một AI hạt giống mới được tạo ra. Nhưng gia cường là một phương pháp lựa chọn động lực có tiềm năng cho những con đường khác tới siêu trí tuệ, bao gồm giả lập não bộ, tăng cường sinh học, giao diện não- máy tính và các mạng lưới và tổ chức, với khả năng xây dựng hệ thống từ những hạt nhân quy chuẩn (những con người bình thường) đã sẵn hàm chứa một phép biểu diễn của giá trị nhân bản.

Sự hấp dẫn của cách tiếp cận gia cường có thể tăng theo tỷ lệ thuận với sự thất vọng của chúng ta về những cách tiếp cận khác cho vấn đề kiểm soát. Tạo ra một hệ thống động lực cho một AI hạt giống sao cho nó luôn duy trì độ an toàn tin cậy được và hữu ích trong suốt quá trình tự cải thiện đệ quy, ngay cả khi hệ thống phát triển thành một siêu trí tuệ hoàn thiện, là một yêu cầu cao, đặc biệt nếu chúng ta phải có được giải pháp đúng ngay từ lần thử đầu tiên.

Với tăng cường, ít nhất chúng ta có thể sẽ bắt đầu với một hệ thống có những động lực quen thuộc và giống người.

Ở mặt trái, có thể khó đảm bảo được rằng một hệ thống động lực phức tạp, đã tiến hóa, thiếu trang nhã và chưa được hiểu biết đến nơi đến chốn, như của con người, sẽ không bị sai lỗi khi bộ máy nhận thức của nó phát triển đến một mức độ cực cao. Như đã thảo luận, một quy trình giả lập não bộ không hoàn hảo nhưng giữ được chức năng trí tuệ có thể không bảo toàn được mọi khía cạnh của tính cách. Điều này cũng đúng (mặc dù có thể sẽ ở mức độ thấp hơn) với các biện pháp tăng cường nhận thức sinh học, điều có ảnh hưởng nhỏ tới động lực, và với các biện pháp tăng cường trí tuệ tập thể của các tổ chức và mạng lưới, điều có thể làm thay đổi tiêu cực các động lực xã hội (chẳng hạn như xê dịch thái độ tập thể đối với người ngoài hay với chính những người ủng hộ). Nếu siêu trí tuệ được tạo ra thông qua bất cứ con đường nào nói trên, một nhà tài trợ dự án cũng sẽ thấy rằng những sự đảm bảo về động lực tối thượng của hệ thống trưởng thành là rất khó có được. Một kiến trúc AI được xác định rõ về toán học và tao nhã về nền tảng có thể, với toàn bộ sự khác biệt về hình thái với con người, trở nên “minh bạch” hơn rất nhiều, ngay cả khi của những khía cạnh quan trọng trong chức năng của nó có thể được xác minh về hình thức.

Chung cuộc, dù ta có liệt kê ưu nhược điểm của phương pháp gia cường ra sao, lựa chọn có nên trông cậy vào nó hay không có thể phải do ép buộc. Nếu siêu trí tuệ được tạo ra đầu tiên qua con đường trí tuệ nhân tạo, phương pháp gia cường sẽ không áp dụng được. Ngược lại, nếu siêu trí tuệ được tạo ra lần đầu qua con đường phi AI nào đó, nhiều phương pháp lựa chọn động lực khác có thể

không áp dụng được. Tuy vậy, các quan điểm về mức độ thành công của phương pháp gia cường vẫn có tính phù hợp chiến lược chừng nào chúng ta còn có cơ hội tác động đến việc công nghệ nào sẽ tạo ra siêu trí tuệ đầu tiên.

Tóm tắt

Có thể làm một tóm tắt nhanh gọn trước khi chúng ta khép lại chương này. Chúng ta đã phân biệt hai lớp phương pháp lớn để xử lý vấn đề đại diện ở trung tâm của an toàn AI: kiểm soát năng lực và lựa chọn động lực. Tóm tắt được đưa ra trong bảng 10.

Mỗi phương pháp có những điểm yếu tiềm tàng riêng của mình và đại diện cho những độ khó khác nhau khi triển khai. Có thể chúng ta sẽ đánh giá chúng từ tốt nhất tới tệ nhất rồi chọn thứ tốt nhất. Nhưng đó sẽ là quá đơn giản. Một số phương pháp có thể được sử dụng tổ hợp trong khi những phương pháp khác lại dùng riêng. Thậm chí một phương pháp tương đối kém an toàn có thể được khuyến nghị nếu có thể dễ dàng sử dụng được như một biện pháp bổ trợ, trong khi một phương pháp mạnh có thể là không hấp dẫn nếu ngăn cản việc sử dụng những biện pháp an toàn mong muốn khác.

Bảng 10 Các phương pháp kiểm soát

Kiểm soát năng lực

Các phương pháp đóng hộp	Hệ thống được cách ly sao cho nó chỉ có thể tác động đến thế giới bên ngoài qua một số kênh hạn chế, đã được phê duyệt trước. Bao gồm các phương pháp cách ly vật lý và thông tin.
Các phương pháp thưởng	Hệ thống được đặt trong một môi trường cung cấp những phần thưởng thích hợp. Đó có thể bao gồm cả việc hòa nhập xã hội vào một thế giới của những thực thể mạnh mẽ tương đương. Một phương án khác là sử dụng các thẻ phần thưởng (mã hóa). "Giám hãm vị nhân" cũng là một khả năng rất quan trọng nhưng là khả năng bao gồm những bí quyết bí truyền.
Úc chế phát triển	Những hạn chế được đặt lên năng lực nhận thức của hệ thống hoặc năng lực tác động đến những quá trình nội bộ quan trọng của nó.
Dây bẫy	Các bài đo chẩn đoán được thực hiện trên hệ thống (có thể nó không biết) và một cơ chế để tắt hệ thống nếu phát hiện bất thường.

Lựa chọn động lực

Đặc tả trực tiếp	Hệ thống được trang bị một hệ thống động lực nào đó được xác định trực tiếp, có thể mang tính hệ quả hay tuân theo một bộ quy tắc.
Tự hạn định	Một hệ thống động lực được thiết kế để hạn chế cao độ phạm vi của các tham vọng và hoạt động của phần mềm tác tử.
Quy chuẩn gián tiếp	Quy chuẩn gián tiếp có thể bao gồm các nguyên tắc theo quy tắc hay hệ quả, nhưng nó được phân biệt bởi việc sử dụng một cách tiếp cận gián tiếp để xác định các quy tắc cần tuân theo hay những giá trị phải theo đuổi.
Gia cường	Ai đó khởi đầu với một hệ thống đã có sẵn những động lực nhân bản hoặc nhân từ đáng kể và nâng cao năng lực nhận thức của nó để biến nó thành một siêu trí tuệ.

Vì vậy, chúng ta cần cân nhắc những sự kết hợp khả thi. Cụ thể, cần quan tâm đến việc sẽ xây dựng loại hệ thống nào, và những phương pháp kiểm soát nào áp dụng được cho từng loại. Đó là chủ đề của chương tiếp theo.

Chương 10

Tiên tri, thần đèn, toàn năng, công cụ

Có người nói: “Chỉ cần xây dựng một hệ thống hỏi-đáp!” hay “Chỉ cần xây dựng một AI như một công cụ thay vì một tác tử phần mềm!” Nhưng những gợi ý này không giải tỏa được các quan ngại về an toàn, và thực tế đó là một câu hỏi không tầm thường về việc kiểu hệ thống nào mang lại những viễn cảnh an toàn tốt nhất. Chúng ta xem xét bốn loại hay “cấp” (tiên tri, thần thánh, toàn năng và công cụ) và giải thích mức độ liên quan giữa bốn cấp này.¹ Mỗi loại cung cấp các tập ưu điểm và nhược điểm khác nhau trong cuộc chinh phục giải quyết vấn đề kiểm soát.

AI Tiên tri

AI tiên tri là một hệ thống hỏi đáp. Nó có thể tiếp nhận các câu hỏi bằng ngôn ngữ tự nhiên và đưa ra câu trả lời dưới dạng văn bản. Một AI tiên tri chỉ nhận các câu hỏi có/không có thể đưa ra dự đoán tốt nhất chỉ bằng một bit, hoặc có thể bổ sung vài bit để thể hiện mức độ tự tin của nó. Một AI tiên tri nhận các câu hỏi mở có thể cần thêm một số tiêu chí để đánh giá những câu trả lời đúng về lượng thông tin và tính phù hợp.² Trong bất cứ trường hợp nào, xây dựng một AI tiên tri có đầy đủ khả năng tư duy chung để trả lời các câu hỏi bằng ngôn ngữ tự nhiên là một vấn đề AI toàn diện. Nếu có thể làm việc đó,

chúng ta cũng có thể chế tạo một AI có năng lực thích hợp để hiểu được ý định và lời nói của con người.

Các AI tiên tri sở hữu siêu trí tuệ trong các ngành hẹp cũng khả thi. Chẳng hạn, chúng ta có thể tạo ra một AI tiên tri-toán học chỉ chấp nhận các truy vấn bằng ngôn ngữ hình thức và rất giỏi trả lời các câu hỏi kiểu này (như, có thể lập tức giải quyết bất kỳ bài toán nào được thể hiện bằng hình thức mà giới toán học chuyên nghiệp phải loay hoay cả thế kỷ). Một AI tiên tri-toán học như vậy có thể tạo thành bàn đạp hướng tới siêu trí tuệ tổng quát.

Đã có các AI tiên tri với siêu trí tuệ trong những ngành siêu hẹp. Máy tính bỏ túi có thể được xem như một AI tiên tri ngành hẹp cho những câu hỏi số học cơ bản; một công cụ tìm kiếm trên Internet có thể được xem như một đại diện rất hạn chế của AI tiên tri với một lĩnh vực bao gồm một phần đáng kể kiến thức khai báo phổ quát của con người. Những AI tiên tri ngành hẹp này là công cụ chứ chưa hẳn là tác tử (chúng ta sẽ sớm nói thêm về AI công cụ). Trong những phần tiếp theo, thuật ngữ “tiên tri” sẽ tham chiếu đến các hệ thống hỏi đáp có siêu trí tuệ tổng quát, trừ khi khái niệm khác được đưa ra.

Để khiến một siêu trí tuệ tổng quát đóng vai trò như một AI tiên tri, chúng ta có thể áp dụng cả hai biện pháp kiểm soát năng lực và lựa chọn động lực. Lựa chọn động lực cho một AI tiên tri có thể dễ dàng hơn so với những cấp khác của siêu trí tuệ, vì mục tiêu tối thượng của AI này tương đối đơn giản. Chúng ta có thể muốn AI tiên tri đưa ra những câu trả lời chân thực, không mang tính thao túng, hoặc giới hạn tác động của nó lên thế giới. Khi áp dụng một phương pháp tự hạn định, chúng ta có thể yêu cầu AI tiên tri chỉ sử dụng những tài nguyên đã được chỉ định để đưa ra câu trả lời. Ví dụ, chúng ta có thể

quy định rằng nó phải trả lời dựa trên nền tảng khối thông tin đã được tải sẵn, ví dụ như một trạng thái đã được lưu trữ của mạng Internet, và chỉ được dùng không nhiều hơn một số lượng bước tính toán cố định nào đó.³

Để tránh khuyến khích AI tiên tri “điều khiển” chúng ta nhằm nhận được những câu hỏi dễ hơn – điều có thể xảy ra nếu chúng ta đặt cho nó mục tiêu tối đa hóa độ chính xác khi trả lời mọi câu hỏi – chúng ta có thể đặt cho nó mục tiêu chỉ trả lời một câu hỏi và tự tắt ngay sau khi trả lời xong. Câu hỏi có thể được tải sẵn vào bộ nhớ của nó trước khi chương trình chạy. Để hỏi câu thứ hai, chúng ta có thể phải khởi động lại máy và chạy cùng một chương trình với một câu hỏi khác đã được nạp sẵn trong bộ nhớ.

Những thách thức khó thấy và nguy hiểm xuất hiện ngay cả trong khi xác định hệ thống động lực tương đối đơn giản nhằm thúc đẩy một AI tiên tri. Giả sử, chúng ta tìm ra được một lời giải thích nào đó về ý nghĩa của việc “giảm thiểu tác động của mình lên thế giới, tùy thuộc vào việc đạt được vài kết quả nào đó”, hoặc “chỉ sử dụng những tài nguyên được chỉ định để trả lời” đối với AI. Chuyện gì xảy ra nếu AI, trong quá trình phát triển trí tuệ của mình, trải qua một sự kiện giống như cách mạng khoa học, gắn với một thay đổi trong bản thể học cơ bản của nó?⁴ Trước hết, có thể ban đầu chúng ta đã giải thích “tác động” và “tài nguyên được chỉ định” bằng bản thể riêng của chúng ta (giả định sự tồn tại của các thực thể vật lý khác nhau như các máy tính). Nhưng cũng giống như chúng ta đã bỏ qua các phạm trù bản thể học khác được các nhà khoa học trong quá khứ coi là đương nhiên (ví dụ “phlogiston” – nhiệt tố, “élan vital” – sức sống, và “tính đồng thời tuyệt đối”), một AI siêu trí tuệ cũng có thể khám phá ra

rằng một số phạm trù hiện tại của chúng ta đã dựa trên những hiểu lầm căn bản. Hệ thống mục tiêu của một AI đang trải qua một cuộc khủng hoảng bản thể học phải đủ mềm dẻo để tiếp nhận “tinh thần” của nội dung mục tiêu gốc và thoải mái chuyển nó thành một “khóa” mới.

Tuy khiến một AI tiên tri trở nên an toàn nhờ phương pháp lựa chọn động lực có thể vô cùng khó khăn, nhưng dù sao vẫn dễ dàng hơn làm thế với một AI chu du khắp thế giới để theo đuổi một mục tiêu phức tạp. Đó là luận điểm ủng hộ việc siêu trí tuệ đầu tiên là một AI tiên tri. Việc này cũng được hỗ trợ bởi một luận điểm khác, cho rằng phương pháp kiểm soát năng lực rất dễ áp dụng cho loại AI này. Tất cả các phương pháp đóng hộp tiêu chuẩn đều áp dụng được ở đây. Ví dụ, hãy xem xét rủi ro xảy ra một AI tiên tri trả lời các câu hỏi không có độ chính xác tối đa mà theo cách nào đó nhằm ngấm ngầm thao túng, khiến chúng ta thúc đẩy một kế hoạch bí mật của riêng nó. Chúng ta có thể giảm nhẹ đôi chút mối đe dọa bằng cách tạo ra AI nhiều tiên tri với mã và cơ sở thông tin hơi khác nhau một chút. Sau đó, câu trả lời của các AI tiên tri khác nhau có thể được so sánh bằng một cơ chế đơn giản, và chỉ đưa ra câu trả lời cho con người khi đã đạt được sự thống nhất. Trong khi cần phải giả định rằng bất cứ AI tiên tri siêu trí tuệ tổng quát nào cũng biết đến sự tồn tại của những AI tiên tri khác (có thể suy ra điều này từ hiểu biết của mình về xã hội loài người), chúng ta vẫn có thể ẩn đi một số chi tiết về thực thi của mỗi AI tiên tri với các AI tiên tri khác. Nếu các AI này không thể giao tiếp, chúng có thể sẽ khó phối hợp để trả lời các câu hỏi của chúng ta một cách thao túng. Có nhiều cách làm lệch lạc sự thật, và các AI tiên tri có thể không nhất trí được sự lệch lạc nào

là hấp dẫn nhất – vì bản thân sự thật là một điểm Schelling (hay tiêu điểm, điểm đồng thuận khi không có sự giao tiếp). Như vậy nếu các tiên tri đạt được đồng thuận, đó có thể là dấu hiệu cho thấy chúng đã đưa ra những câu trả lời đúng.⁵

Một AI tiên tri có thể đáng tin cậy một cách lý tưởng theo nghĩa chúng ta có thể an toàn giả định rằng nó sẽ luôn đưa ra câu trả lời đúng nhất trong phạm vi khả năng của mình. Nhưng ngay cả một AI tiên tri không đáng tin cậy cũng có thể có ích. Chúng ta có thể hỏi AI tiên tri này những câu hỏi khó trả lời nhưng để xác minh xem câu trả lời được đưa ra có chính xác hay không. Đây là dạng thức của nhiều bài toán. Nếu chưa rõ liệu một mệnh đề toán học là đúng hay sai, chúng ta có thể yêu cầu một AI tiên tri chứng minh hay phản chứng mệnh đề đó. Việc tìm cách chứng minh có thể cần đến sự thấu hiểu và sáng tạo ngoài tầm của chúng ta, nhưng kiểm chứng tính đúng đắn của một phép chứng minh nổi tiếng nào đó có thể thực hiện được bằng một quy trình đơn giản.

Nếu việc kiểm chứng các câu trả lời là quá đắt đỏ (thường là các chủ đề nằm ngoài phạm vi logic và toán học), chúng ta có thể lựa chọn ngẫu nhiên một tập hợp nhỏ các câu trả lời của AI tiên tri để kiểm chứng. Nếu toàn bộ tập hợp này là đúng, chúng ta có thể ấn định một xác suất cao cho việc hầu hết các câu trả lời khác là chính xác. Thủ thuật này có thể giảm đáng kể chi phí kiểm chứng đắt đỏ với từng câu trả lời. (Không may, nó không thể đưa ra những câu trả lời đáng tin cậy mà chúng ta không thể kiểm chứng, vì một AI tiên tri đang lừa dối có thể chọn chỉ trả lời đúng những câu hỏi mà nó tin rằng chúng ta có thể kiểm chứng câu trả lời).

Có thể có những vấn đề quan trọng mà trong đó, chúng ta được lợi từ một con trỏ mang tính đoán định tới câu trả lời đúng (hoặc tới một phương pháp xác định câu trả lời đúng) thậm chí nếu chúng ta phải tích cực nghi ngờ nguồn gốc. Ví dụ, ta có thể yêu cầu lời giải cho những vấn đề kỹ thuật và triết học khác nhau phát sinh trong khi tìm cách phát triển những phương pháp lựa chọn động lực tiên tiến hơn. Nếu sở hữu một thiết kế AI được đề xuất và được giả định là an toàn, chúng ta có thể hỏi một AI tiên tri xem nó có thể xác định được sai sót nào trong thiết kế, và liệu nó có thể giải thích những thiếu sót đó cho chúng ta chỉ bằng 20 từ hoặc ít hơn hay không. Tuy nhiên chúng ta cần phải thận trọng và kiểm chế để không đặt quá nhiều câu hỏi – và không để cho bản thân tiêu hóa quá nhiều chi tiết của các câu trả lời được đưa ra cho những câu hỏi của mình – tạo cơ hội để các AI tiên tri không đáng tin cậy tác động vào tâm lý của mình (bằng cách sử dụng các thông điệp có vẻ khả thi nhưng ngầm mang tính thao túng). Có thể một AI với siêu năng lực thao túng xã hội không cần quá nhiều tương tác để khiến chúng ta phục tùng ý chí của nó.

Thậm chí nếu bản thân AI tiên tri làm việc đúng như dự kiến, vẫn có rủi ro về việc nó có thể bị sử dụng sai mục đích. Một chiều kích rõ ràng của vấn đề này là một AI tiên tri có thể là nguồn quyền năng khổng lồ, có thể tạo ra ưu thế chiến lược quyết định cho người vận hành nó. Quyền năng này có thể không hợp lệ và có thể không được sử dụng vì lợi ích chung. Một chiều kích tinh tế nhưng không kém phần quan trọng khác là việc sử dụng một AI tiên tri có thể vô cùng nguy hiểm cho chính người vận hành. Những lo lắng tương tự (có liên quan đến các vấn đề triết lý cũng như kỹ thuật) cũng xuất hiện cho những cấp giả thuyết khác của siêu trí tuệ. Chúng ta sẽ khám

phá chúng kỹ hơn trong Chương 13. Ở đây, ta chỉ cần lưu ý rằng quy trình xác định câu hỏi nào được hỏi, theo trình tự nào và các câu trả lời được báo cáo và phân bổ như thế nào có thể có tầm quan trọng đặc biệt. Ta cũng có thể xem xét liệu có nên tìm cách chế tạo AI tiên tri sao cho nó sẽ từ chối trả lời bất cứ câu hỏi nào trong những trường hợp mà nó dự báo rằng câu trả lời của mình có thể đem lại những hậu quả được coi là có tính thảm họa theo một số tiêu chí “thô nhưng thật”.

AI thần đèn và AI toàn năng

AI thần đèn là một hệ thống thực hiện mệnh lệnh: nó nhận một mệnh lệnh cấp cao, thực hiện nó, rồi dừng lại để chờ lệnh tiếp theo.⁶ AI toàn năng là một hệ thống có thể thoải mái hoạt động trên thế giới để theo đuổi những mục tiêu rộng lớn và có thể rất dài hạn. Mặc dù các loại này có vẻ như là các khuôn mẫu hoàn toàn khác nhau về bản chất và vai trò của siêu trí tuệ, nhưng sự khác biệt này không quá sâu sắc như những gì thoát thấy.

Với một AI thần đèn, ta đã hy sinh đặc tính hấp dẫn nhất của một AI tiên tri: cơ hội sử dụng các phương pháp đóng hộp. Trong khi ta có thể xem xét việc tạo ra một AI thần đèn bị giam hãm về mặt vật lý, như một AI thần đèn chỉ có thể tạo ra các đối tượng bên trong một không gian được chỉ định – một không gian có thể bị bọc kín bởi một bức tường được gia cố hoặc một hàng rào gắn kíp nổ sẵn sàng kích hoạt nếu vùng cách ly bị xâm phạm – có thể khó tin tưởng được vào sự an toàn của bất cứ phương pháp cách ly vật lý nào tương tự khi chống lại một siêu trí tuệ được trang bị tay máy linh hoạt và vật liệu xây dựng.

Thậm chí nếu bằng cách nào đó có thể đảm bảo được một nơi cách ly an toàn như đối với một AI tiên tri thì cũng chưa rõ chúng ta sẽ được lợi bao nhiêu khi cho siêu trí tuệ tiếp cận trực tiếp đến những cánh tay máy so với việc yêu cầu nó đưa ra các bản vẽ mà chúng ta có thể kiểm định rồi sử dụng để tự mình tạo ra kết quả. Lợi ích có được về tốc độ và sự thuận tiện từ việc bỏ qua những người trung gian có vẻ như không hề xứng đáng với việc bỏ qua các phương pháp đóng hộp mạnh hơn sẵn có để cách ly một AI tiên tri.

Nếu ai đó muốn tạo ra một AI thần đèn, có thể nên chế tạo nó để tuân theo dự tính đằng sau mệnh lệnh thay vì nghĩa đen của nó, vì một AI thần đèn chỉ hiểu nghĩa đen (đủ siêu trí tuệ tới mức có được ưu thế chiến lược quyết định) có thể có khả năng tiêu diệt người dùng và cả nhân loại trong lần sử dụng đầu tiên vì những lý do được nêu trong phần nói về các chế độ thất bại ác tính (Chương 8). Nói rộng hơn, một điều có vẻ quan trọng là AI thần đèn tìm kiếm một cách giải thích khoan dung – có thể được con người coi là hợp lý – của mệnh lệnh, và rằng AI thần đèn có thể được khuyến khích bằng cách thực hiện mệnh lệnh theo cách giải thích như vậy thay vì nghĩa đen. Một AI thần đèn lý tưởng sẽ là một “siêu quản gia” thay vì một thiên tài tự kỷ.

Tuy nhiên một AI thần đèn có bản chất siêu quản gia có thể không cách xa lắm với việc được chấp nhận thành thành viên của đẳng cấp các AI toàn năng. Để so sánh, hãy xem xét ý tưởng chế tạo một AI toàn năng với mục tiêu tối thượng là phục tùng tinh thần của các mệnh lệnh mà chúng ta đưa ra rồi chúng ta chế tạo một AI thần đèn thay vì một AI toàn năng. Một AI toàn năng như vậy có thể bắt chước một AI thần đèn. Là siêu trí tuệ, AI toàn năng này có thể sẽ

đoán khá đúng các mệnh lệnh chúng ta sẽ ra cho một AI thần đèn (và nó luôn có thể hỏi chúng ta nếu điều đó giúp nó ra quyết định). Liệu có khác biệt nào thực sự quan trọng giữa một AI thần đèn và một AI toàn năng? Hoặc nếu khai thác sự khác biệt từ phía bên kia, hãy chú ý rằng một AI thần đèn siêu trí tuệ tương tự như vậy cũng có thể dự đoán được chúng ta sẽ ra cho nó những mệnh lệnh gì: vậy chờ lệnh ban ra rồi mới thực hiện thì sẽ được lợi điều gì?

Ta có thể nghĩ rằng một ưu điểm lớn của AI thần đèn so với một AI toàn năng là nếu có gì đó sai thì chúng ta có thể ra cho AI thần đèn một mệnh lệnh mới để dừng hay đảo ngược hệ quả của những hành động trước đó, bởi vì một AI toàn năng sẽ cứ tiếp tục thực thi bất chấp sự phản đối của chúng ta. Nhưng ưu thế an toàn rõ ràng này với AI thần đèn chủ yếu chỉ là ảo tưởng. Nút “dừng” và “khôi phục” trên một AI thần đèn chỉ có tác dụng đối với các chế độ thất bại lành tính: trong trường hợp thất bại ác tính – trường hợp mà trong đó, chẳng hạn, việc thực thi mệnh lệnh hiện thời đã trở thành mục tiêu tối thượng của AI thần đèn – thì nó sẽ đơn giản gạt bỏ mọi nỗ lực sau đó nhằm hóa giải mệnh lệnh ban đầu.⁷

Lựa chọn khả thể là chế tạo một AI thần đèn có thể tự động giới thiệu cho người sử dụng về những khía cạnh nổi bật của các kết quả khả thể từ một mệnh lệnh được đề xuất, và hỏi để khẳng định trước khi thực hiện. Một hệ thống như vậy có thể được tham chiếu đến như một AI thần đèn có bản xem trước. Nhưng nếu có thể làm điều đó cho một AI thần đèn, thì cũng có thể được thực hiện với một AI toàn năng. Như vậy, một lần nữa ta vẫn chưa thấy rõ được yếu tố khác biệt giữa một AI thần đèn và một AI toàn năng. (Giả sử chức năng xem trước có thể được tạo ra thì việc liệu có thể sử dụng nó

hay không, cùng với cách thức sử dụng, lại ít hiển nhiên hơn chúng ta tưởng, bất kể sức hấp dẫn mạnh mẽ của việc có thể thoáng nhìn kết quả trước khi cam kết biến nó thành thực tiễn không thể đảo ngược. Chúng ta sẽ quay lại với vấn đề này sau.)

Khả năng một cấp AI này bắt chước một cấp AI khác cũng có thể mở rộng sang các AI tiên tri. Một AI thần đèn có thể bị buộc hành động như một AI tiên tri nếu những mệnh lệnh duy nhất mà chúng ta ra cho nó là phải trả lời những câu hỏi nhất định. Một AI tiên tri, tới lượt mình, có thể được làm ra để thay thế cho một AI thần đèn nếu như chúng ta hỏi AI tiên tri cách dễ nhất để thực hiện những mệnh lệnh nhất định nào đó là gì. AI tiên tri có thể cho chúng ta những chỉ dẫn từng bước để đạt được cùng những kết quả như một AI thần đèn có thể đạt được, hoặc thậm chí nó có thể đưa ra mã nguồn cho một AI thần đèn.⁸ Những quan điểm tương tự có thể được đưa ra liên quan đến mối quan hệ giữa một AI tiên tri và một AI toàn năng.

Sự khác biệt thực sự giữa ba cấp này do đó không nằm trong những năng lực tối thượng của chúng. Thay vào đó, sự khác biệt được quy về những cách tiếp cận khác nhau tới vấn đề kiểm soát. Mỗi cấp tương ứng với một tập hợp các biện pháp phòng ngừa an toàn khác biệt. Tính năng nổi bật nhất của một AI tiên tri là có thể giam giữ được nó. Một AI thần đèn khó bị giam giữ hơn, nhưng ít nhất có thể áp dụng biện pháp thuần hóa. Một AI toàn năng không thể bị giam cầm hay kiểm soát bằng cách tiếp cận thuần hóa.

Nếu như đó chỉ là những nhân tố liên quan duy nhất thì thứ tự của mức độ mong muốn có vẻ như rõ ràng: một AI tiên tri sẽ an toàn hơn một AI thần đèn, tới lượt mình lại an toàn hơn AI toàn năng; và mọi khác biệt ban đầu liên quan đến sự thuận tiện và tốc độ vận hành

có thể tương đối nhỏ và dễ dàng bị chèn ép bởi những lợi ích thu được từ tính an toàn khi chế tạo AI tiên tri. Tuy nhiên còn có những yếu tố khác cần được tính đến. Khi chọn lựa giữa các cấp độ ta cần xem xét không chỉ những mối nguy hiểm gây ra bởi chính bản thân hệ thống mà còn những mối nguy hiểm xuất hiện từ cách sử dụng nó. Một AI thần đèn sẽ cho người kiểm soát được nó quyền năng cực lớn, nhưng một AI tiên tri cũng vậy.⁹ Ngược lại, một AI toàn năng có thể được chế tạo sao cho không chịu bất cứ ảnh hưởng đặc biệt của bất cứ cá nhân hay nhóm người nào tới kết quả, và như vậy nó sẽ chống lại mọi nỗ lực sửa đổi hoặc làm sai lệch kế hoạch ban đầu của mình. Hơn nữa, nếu như động lực của một AI toàn năng được xác định bằng “quy chuẩn gián tiếp” (một khái niệm sẽ được thảo luận thêm trong Chương 13), nó có thể được sử dụng để tạo ra kết quả được xác định trừu tượng, chẳng hạn như “bất cứ điều gì tối công bằng và đúng đắn về đạo đức” – mà không ai biết trước được điều đó có hệ quả ra sao. Nó có thể tạo ra một tình huống tương tự như một “bức màn vô thức” (veil of ignorance) của Rawls.¹⁰ Bối cảnh này có thể tạo thuận lợi cho việc có được sự đồng thuận, giúp ngăn chặn xung đột và thúc đẩy kết quả công bằng hơn.

Một ý khác chống lại một số loại AI tiên tri và thần đèn là có thể phát sinh rủi ro trong khi thiết kế một siêu trí tuệ với một mục tiêu tối thượng không hoàn toàn phù hợp với kết quả mà cuối cùng chúng ta muốn đạt được. Ví dụ, khi dùng một động lực tự hạn định để khiến một siêu trí tuệ muốn giảm thiểu một số tác động của mình lên thế giới, chúng ta có thể từ đó tạo ra một hệ thống có cách đánh giá các kết quả khác với người tài trợ. Điều tương tự sẽ xảy ra khi xây dựng một AI đặt giá trị đặc biệt cao vào việc trả lời đúng các câu hỏi hay

trung thành tuân thủ từng mệnh lệnh. Như vậy, nếu được quan tâm đúng mức, điều này sẽ không gây ra bất cứ vấn đề gì: có thể sẽ có thỏa thuận thích hợp giữa hai cách đánh giá – ít nhất cho tới khi chúng thuộc về những thế giới khả thể tương đối dễ hiện thực hóa – rằng những kết quả nào tốt theo tiêu chuẩn của AI cũng tốt theo tiêu chuẩn của con người. Nhưng ta có thể ủng hộ nguyên tắc thiết kế cho rằng ngay cả việc chỉ đưa vào đôi chút bất hòa giữa các mục tiêu của AI và của chúng ta cũng không phải khôn ngoan. (Mối quan ngại tương tự tất nhiên cũng xuất hiện với việc đặt ra cho các AI toàn năng những mục tiêu không hoàn toàn hài hòa với mục tiêu của con người).

AI công cụ

Một gợi ý cho rằng chúng ta có thể chế tạo siêu trí tuệ như một công cụ chứ không phải một tác tử.¹¹ Ý tưởng này xem ra xuất phát từ quan sát cho rằng phần mềm thông thường, được sử dụng trong vô vàn các ứng dụng, không gây nên bất cứ quan ngại về an toàn nào dù chỉ tương đồng chút ít với những gì được thảo luận trong cuốn sách này. Vì sao ta không thể tạo ra “AI công cụ” giống như các phần mềm – chẳng hạn như một hệ thống điều khiển bay, hay một trợ lý ảo – nhưng linh hoạt và có năng lực hơn? Vì sao phải chế tạo một siêu trí tuệ có ý chí riêng? Theo cách nghĩ này, mô hình tác tử là hoàn toàn sai lầm. Thay vì tạo ra một AI có niềm tin và mong muốn, và hành động như một cá nhân nhân tạo, chúng ta nên hướng tới mục tiêu xây dựng một phần mềm chỉ thực thi những gì nó được lập trình.

Tuy nhiên, ý tưởng tạo ra một phần mềm chỉ “thực thi những gì nó được lập trình” lại không đơn giản như vậy nếu sản phẩm được

tạo ra là một trí tuệ tổng quát mạnh mẽ. Tất nhiên, chúng ta thường sẽ nghĩ rằng tất cả phần mềm chỉ thực thi những gì được lập trình: hành vi được xác định bằng toán học thông qua hệ thống mã. Nhưng điều này đúng cho mọi cấp trí tuệ máy, cho dù là AI công cụ hay không. Nếu thay vào đó, việc “thực thi những gì nó được lập trình” lại có nghĩa là phần mềm hành xử như các lập trình viên mong muốn, thì đó là một tiêu chuẩn mà phần mềm thông thường ít khi đạt được.

Vì phần mềm hiện tại có năng lực hạn chế (so với những phần mềm của siêu trí tuệ máy) nên hệ quả của những thất bại này là có thể quản lý được, từ những hệ quả không đáng kể cho tới rất đắt giá, nhưng không trong trường hợp nào tạo ra các mối đe dọa diệt vong.¹²

Tuy nhiên, nếu thiếu năng lực (chứ không phải thiếu độ tin cậy) mới là thứ khiến cho phần mềm thông thường trở nên an toàn khi bàn về vấn đề diệt vong, thì cách để khiến phần mềm này trở thành hình mẫu cho một siêu trí tuệ an toàn vẫn còn chưa rõ. Có thể có suy nghĩ cho rằng bằng cách mở rộng phạm vi tác vụ mà phần mềm thông thường thực hiện, ta có thể loại bỏ nhu cầu về trí tuệ nhân tạo tổng hợp. Nhưng phạm vi và sự đa dạng của các tác vụ mà một trí tuệ tổng quát có thể thực hiện để đem lại lợi nhuận trong một nền kinh tế hiện đại là rất lớn, và có thể việc tạo ra phần mềm chuyên dụng để xử lý tất cả những tác vụ này là không khả thi. Dù làm được đi nữa, thì dự án như vậy có thể tốn rất nhiều thời gian. Trước khi nó hoàn thành, bản chất của một số tác vụ có thể đã thay đổi và những tác vụ mới có thể phát sinh. Việc sở hữu phần mềm có thể tự học để thực hiện tác vụ mới và tìm ra những tác vụ mới cần thực hiện là một lợi thế lớn. Nhưng điều này đòi hỏi phần mềm phải có khả năng học,

lý luận và lập kế hoạch với năng lực mạnh mẽ và bao phủ nhiều lĩnh vực. Nói cách khác, nó cần trí tuệ tổng quát.

Đặc biệt phù hợp với các mục tiêu của chúng ta lại chính là tác vụ phát triển phần mềm. Tự động hóa được công việc này sẽ đem lại những ưu thế thực tế vô cùng to lớn. Mặc dù vậy, năng lực tự cải thiện nhanh chóng chính là đặc tính thiết yếu cho phép một AI hạt giống chầm ngòi cho một sự bùng nổ trí tuệ.

Nếu trí tuệ tổng quát là bắt buộc, liệu còn cách nào khác để diễn dịch ý tưởng AI công cụ nhằm bảo toàn được phẩm chất thụ động giúp chúng ta yên tâm của một công cụ nhàm chán? Liệu chúng ta có thể có một trí tuệ tổng quát không phải là tác tử? Về mặt trực giác, không phải chỉ năng lực hạn chế của phần mềm thông thường khiến nó an toàn mà cả sự thiếu tham vọng của nó cũng vậy. Không có một trình con nào trong Excel âm thầm muốn chiếm lĩnh thế giới vì nó không đủ thông minh để tìm ra cách. Ứng dụng bảng tính không “muốn” bất kỳ thứ gì; nó chỉ mù quáng thực hiện các chỉ dẫn trong chương trình. (Ta có thể tự hỏi) điều gì ngăn cản việc tạo ra một ứng dụng có trí tuệ tổng quát cao hơn cùng loại, chẳng hạn như một AI tiên tri có thể phản hồi khi được gán một mục tiêu bằng một kế hoạch thực hiện mục tiêu đó (rất giống với cách Excel phản hồi một cột số bằng cách tính tổng – mà không thể hiện bất kỳ sự “quan tâm” nào về đầu ra hoặc cách thức sử dụng của tổng)?

Phương pháp viết phần mềm truyền thống đòi hỏi lập trình viên hiểu được tác vụ cần thực hiện đủ chi tiết để tạo ra một quá trình giải quyết rõ ràng, bao gồm một chuỗi các bước được định nghĩa rõ bằng toán học và có thể thể hiện bằng mã.¹³ (Trên thực tế, các kỹ sư phần mềm trông cậy vào các thư viện mã chứa những hành vi hữu

ích có thể vận dụng mà không cần hiểu cách thức các hành vi này được triển khai. Nhưng mã đó ban đầu vẫn được khởi tạo bởi những lập trình viên có hiểu biết chi tiết về công việc của mình.) Cách tiếp cận này phù hợp để giải quyết các tác vụ đã được hiểu rõ và được dùng để viết ra hầu hết phần mềm đang sử dụng hiện nay. Tuy nhiên, chúng ta không thể áp dụng nó vì không ai biết chính xác làm thế nào để giải quyết tất cả các tác vụ cần hoàn thành. Đây chính là lúc các kỹ thuật trí tuệ nhân tạo trở nên phù hợp. Trong những ứng dụng hẹp, học máy có thể chỉ được dùng để tinh chỉnh một số ít tham số trong một chương trình lớn do con người thiết kế. Chẳng hạn, một bộ lọc thư rác có thể được huấn luyện bằng một tập thư điện tử đã phân loại bằng tay trong một quá trình làm thay đổi các trọng số mà thuật toán phân loại gán cho các tính năng chẩn đoán khác nhau. Trong một ứng dụng tham vọng hơn, bộ phân loại có thể được xây dựng để tự khám phá và kiểm định được hiệu lực của những tính năng mới trong một môi trường thay đổi. Một bộ lọc thư rác phức tạp hơn nữa có thể được trang bị khả năng suy luận về những cân nhắc mà người dùng phải đối mặt, hoặc về nội dung của các bức thư mà nó đang phân loại. Trong cả hai trường hợp này, lập trình viên đều không cần biết cách tốt nhất để phân biệt giữa thư rác và thư thường mà chỉ cần biết cách thiết lập một thuật toán có thể tự cải thiện được hiệu năng nhờ học tập, khám phá hay suy luận.

Cùng với những tiến bộ trong trí tuệ nhân tạo, lập trình viên có thể giảm tải nhiều hơn những công việc trí óc cần thiết để biết được cách thực hiện một tác vụ đã cho. Trong một trường hợp đặc biệt, lập trình viên có thể chỉ cần xác định một tiêu chí chính thức cho một kết quả thành công và để AI tự tìm giải pháp. Để hướng dẫn việc tìm

kiểm của mình, AI có thể sử dụng một bộ công cụ tự nghiệm và các phương pháp khác để khám phá cấu trúc trong không gian các giải pháp khả thể. Nó sẽ tiếp tục tìm kiếm cho đến khi tìm thấy một giải pháp thỏa mãn tiêu chí thành công. AI sau đó có thể tự triển khai giải pháp hoặc (trong trường hợp AI tiên tri) báo cáo cho người dùng về giải pháp.

Những dạng thức sơ đẳng của cách tiếp cận này hiện đang được triển khai khá rộng rãi. Mặc dù vậy, phần mềm sử dụng AI và các kỹ thuật học, dù đã có một số năng lực tìm ra những giải pháp mà các lập trình viên không ngờ tới, vẫn hoạt động như một công cụ cho tất cả các mục tiêu thực tiễn và không gây ra rủi ro diệt vong nào. Chúng ta chỉ bước vào khu vực nguy hiểm khi những phương pháp được sử dụng để tìm kiếm giải pháp trở nên cực kỳ mạnh mẽ và tổng thể: có nghĩa là khi chúng bắt đầu gần với trí tuệ tổng quát – và đặc biệt khi chúng bắt đầu gần với siêu trí tuệ.

Có (ít nhất) hai vùng mà từ đó có thể nảy sinh rắc rối. Thứ nhất, quá trình tìm kiếm của siêu trí tuệ có thể tìm ra một giải pháp không chỉ bất ngờ mà hoàn toàn nằm ngoài dự kiến. Điều đó có thể dẫn đến một trong những kiểu thất bại mà chúng ta đã thảo luận trước đây (“hiện thực hóa sai lệch”, “dư thừa hạ tầng” hoặc “tội ác tâm trí”). Điều này được thể hiện rõ ràng nhất trong trường hợp của một AI toàn năng hoặc thần đèn, những đối tượng trực tiếp triển khai giải pháp mà chúng tìm ra. Nếu vẽ mặt cười kích thích phân tử hay biến cả hành tinh thành kẹp giấy là ý tưởng đầu tiên mà siêu trí tuệ phát hiện ra là đáp ứng được tiêu chí của giải pháp thì chúng ta sẽ có mặt cười và kẹp giấy.¹⁴ Nhưng ngay cả một AI tiên tri chỉ báo cáo giải pháp khi mọi vấn đề khác đều ổn cũng có thể trở thành nguyên nhân

của sự hiện thực hóa sai lệch. Người dùng yêu cầu AI tiên tri lập một kế hoạch để đạt được một kết quả nhất định, hoặc yêu cầu một công nghệ để phục vụ cho một chức năng nhất định; khi người dùng làm theo kế hoạch hoặc tạo ra công nghệ, một sự hiện thực hóa sai lệch có thể nảy sinh giống như trường hợp AI tự triển khai giải pháp.¹⁵

Vùng thứ hai có thể phát sinh rắc rối là quá trình vận hành phần mềm. Nếu những phương pháp mà phần mềm sử dụng để tìm kiếm giải pháp đủ tinh vi thì chúng có thể bao gồm những điều kiện để tự quản lý quá trình tìm kiếm một cách thông minh. Trong trường hợp này, máy chạy phần mềm có thể bắt đầu khác một công cụ và trở nên giống với một tác tử hơn. Do đó, phần mềm có thể khởi đầu bằng cách phát triển một kế hoạch nhằm tìm kiếm giải pháp. Kế hoạch này có thể chỉ ra khu vực nào cần khám phá trước và bằng phương pháp nào, cần thu thập dữ liệu gì và làm thế nào để sử dụng tốt nhất những nguồn tài nguyên tính toán sẵn có. Khi tìm kiếm một kế hoạch đáp ứng được tiêu chí nội bộ của phần mềm (chẳng hạn, tạo ra được xác suất đủ cao cho việc tìm kiếm một giải pháp thỏa mãn tiêu chí do người dùng xác định trong khoảng thời gian được cung cấp), phần mềm có thể gặp một ý tưởng phi chính thống. Ví dụ nó có thể lập ra một kế hoạch bắt đầu bằng việc chiếm hữu những nguồn lực tính toán bổ sung và loại bỏ những đối tượng có thể làm gián đoạn (ví dụ như con người). Những kế hoạch “sáng tạo” như vậy hiển hiện khi năng lực nhận thức của phần mềm đạt đến mức độ đủ cao. Khi phần mềm đưa kế hoạch này vào thực hiện, một thảm họa diệt vong có thể xảy đến.

Như được minh họa bằng các ví dụ trong Khung 9, các quá trình tìm kiếm mở đôi khi hiển lộ những giải pháp khác thường không phù

hợp với con người và nằm ngoài dự kiến, cho dù vẫn ở trong những dạng thức hạn chế hiện nay của chúng. Các quá trình tìm kiếm ngày nay không nguy hại vì chúng quá yếu để có thể tìm ra loại kế hoạch phục vụ cho một chương trình chiếm lĩnh thế giới. Một kế hoạch như vậy có thể bao gồm những bước đi cực kỳ khó khăn, ví dụ như sáng chế ra một công nghệ vũ khí mới đi trước công nghệ hiện tại một vài thế hệ, hoặc thực hiện một chiến dịch tuyên truyền hiệu quả hơn nhiều so với các hình thức truyền thông do bất cứ tuyên truyền viên nào thực hiện. Để có được một cơ hội dù chỉ là thai nghén những ý tưởng đó chứ chưa nói đến việc phát triển chúng để thực sự có hiệu quả, một cỗ máy có lẽ phải cần có khả năng thể hiện thế giới theo cách thức ít nhất cũng phong phú và hiện thực như mô hình thế giới của một người lớn bình thường (mặc dù việc thiếu nhận thức trong một số lĩnh vực có thể được bù đắp bởi kỹ năng bổ sung trong những lĩnh vực khác). Điều này vượt quá tầm với của AI đương đại. Và do sự bùng nổ tổ hợp mà về cơ bản sẽ đánh bại mọi nỗ lực giải quyết các vấn đề lập kế hoạch phức tạp bằng các phương pháp tổng lực (như chúng ta đã thấy trong Chương 1), không thể thực sự vượt qua những hạn chế của các thuật toán đã biết chỉ bằng cách đổ thêm công suất tính toán.¹⁶

Tuy nhiên, một khi các quá trình tìm kiếm hay lập kế hoạch đã trở nên đủ mạnh, chúng có khả năng trở nên nguy hiểm.

Thay vì cho phép hành vi có mục tiêu tương tự như tác tử hình thành tự phát và tình cờ từ việc triển khai các quá trình tìm kiếm mạnh mẽ (bao gồm các quá trình tìm kiếm những kế hoạch công việc nội bộ và các quá trình trực tiếp tìm kiếm các giải pháp thỏa mãn một số chỉ tiêu do người dùng xác định), có thể tốt hơn nếu cố tình tạo ra

các tác tử. Trang bị cho siêu trí tuệ một cấu trúc như tác tử có thể là một cách tăng tính dễ đoán và dễ thấy. Một hệ thống được thiết kế tốt, được chế tạo sao cho có sự phân biệt rõ ràng giữa các niềm tin và các giá trị của nó, có thể sẽ cho phép chúng ta dự báo được điều gì đó về những kết quả mà nó có thể tạo ra. Ngay cả khi chúng ta không thể nhìn thấy trước chính xác hệ thống sẽ có những niềm tin nào, hay rơi vào những tình huống nào, sẽ vẫn có một vùng quen thuộc mà ở đó, chúng ta có thể kiểm tra lại các giá trị cuối cùng của nó và từ đó là cả các tiêu chí mà nó sẽ sử dụng khi chọn lựa những hành động tương lai và đánh giá bất cứ kế hoạch tiềm năng nào.

Khung 9: Những giải pháp lạ lùng xuất phát từ kiểm tìm mù quáng

Ngay cả các quá trình tìm kiếm tiến hóa đơn giản đôi khi cũng đem lại các kết quả vô cùng bất ngờ, hay những giải pháp thỏa mãn được một tiêu chí chính thức do người dùng xác định theo một cách rất khác với so với kỳ vọng hay dự kiến của người dùng.

Lĩnh vực phần cứng tiến hóa cho ta nhiều minh họa về hiện tượng này. Trong lĩnh vực này, một thuật toán tiến hóa tìm kiếm trong không gian các thiết kế phần cứng và kiểm tra tính phù hợp của mỗi thiết kế bằng cách thực sự chế ra nó trên một bảng hoặc bo mạch chủ có thể tái cấu hình nhanh chóng.

Các thiết kế đã tiến hóa thường thể hiện tính kinh tế khá cao. Ví dụ, một quá trình tìm kiếm phát hiện ra một mạch tách tần chạy không cần đồng hồ – một linh kiện thường được xem như cần thiết cho chức năng này. Các nhà nghiên cứu đánh giá rằng mạch tiến hóa này có kích thước nhỏ hơn những gì mà một kỹ sư cần cho công việc khoảng từ 10 đến 100 lần. Mạch này khai thác các tính chất vật lý

của các linh kiện của nó theo cách thức phi chính thống; một vài linh kiện dùng nguồn cần thiết thậm chí còn chưa kết nối với các chân đầu vào và đầu ra! Thay vào đó những linh kiện này được nối qua những thứ thông thường có thể xem như các hiệu ứng phụ không mong muốn như ghép điện từ hoặc phụ tải nguồn.

Một quá trình tìm kiếm khác, có trách nhiệm tạo ra một bộ dao động mà thiếu mất một linh kiện có vẻ như không thể thiếu là tụ điện. Khi thuật toán trình bày giải pháp thành công của nó, các nhà nghiên cứu kiểm tra và đầu tiên kết luận rằng nó “không thể hoạt động được”. Sau khi kiểm tra lại kỹ lưỡng hơn, họ phát hiện ra rằng thuật toán đã tái cấu hình bo mạch chủ không có cảm biến của nó thành một máy thu vô tuyến tạm thời theo kiểu MacGyver, sử dụng các đường nối của bảng mạch in làm ăng ten để bắt tín hiệu do các máy tính cá nhân trong phòng thí nghiệm tình cờ nằm gần đó phát ra. Mạch khuếch đại tín hiệu này để tạo ra tín hiệu dao động đầu ra mong muốn.¹⁷

Trong một thí nghiệm khác, các thuật toán tiến hóa thiết kế các mạch cảm nhận được liệu bo mạch chủ có bị giám sát bởi máy hiển thị dao động hay không hoặc liệu kim loại hàn dán có nối với nguồn điện chung của phòng thí nghiệm hay không.

Những ví dụ này minh họa cách làm thế nào mà một quá trình tìm kiếm mở có thể điều chỉnh mục đích sử dụng của các vật liệu mà nó tiếp cận được để tạo ra những năng lực cảm biến hoàn toàn bất ngờ, theo những cách thức mà tư duy thiết kế thông thường của con người không được trang bị tốt để khai thác hay thậm chí tính đến.

Xu hướng tìm kiếm tiến hóa để “ăn gian” hay tìm những biện pháp phản trực giác để đạt được một mục tiêu cụ thể cũng được thể

hiện cả trong tự nhiên, tuy có thể kém rõ ràng hơn vì vốn chúng ta đã quen thuộc với những đặc trưng sinh học và mặc nhiên coi kết quả của quá trình tiến hóa là bình thường (ngay cả khi không có tiên đoán). Nhưng, việc thiết kế các thí nghiệm về chọn lọc nhân tạo mà trong đó, ta có thể thấy được quá trình tiến hóa hoạt động bên ngoài bối cảnh thông thường của nó vẫn là khả thi. Trong các thí nghiệm này, các nhà nghiên cứu có thể tạo ra những điều kiện ít xảy ra trong tự nhiên và quan sát kết quả.

Ví dụ, trước những năm 1960, các nhà sinh vật học thường hay duy trì quan điểm rằng các loài thú săn mồi hạn chế sự sinh sản của mình để tránh rơi vào bẫy Malthus.¹⁸ Mặc dù chọn lọc cá nhân có thể hành động chống lại cản trở này, nhưng đôi khi người ta cho rằng chọn lọc nhóm sẽ vượt qua những khuyến khích cá nhân để khai thác các cơ hội sinh sản và ưu tiên những phẩm chất có thể hữu ích cho nhóm hay cho cả loài nói chung. Những nghiên cứu phân tích lý thuyết và mô phỏng sau này cho thấy rằng trong khi chọn lọc nhóm là có thể về mặt nguyên tắc, nó có thể vượt qua được chọn lọc cá nhân mạnh mẽ chỉ trong những điều kiện nghiêm ngặt ít khi có được trong tự nhiên.¹⁹ Nhưng những điều kiện này có thể được tạo ra trong phòng thí nghiệm. Khi những con bọ bột mì (*Tribolium castaneum*) được nuôi dưỡng để giảm kích thước quần thể, bằng cách áp dụng tiến hóa nhóm, tiến hóa thực sự đã tạo ra những quần thể nhỏ hơn.²⁰ Tuy nhiên, những biện pháp được sử dụng để đạt được điều đó không chỉ bao gồm những phương thức thích ứng “lành tính” như giảm khả năng sinh sản và kéo dài thời gian phát triển mà một quá trình tìm kiếm tiến hóa nhân hình hóa ngây thơ của loài người có thể kỳ vọng, mà còn cả xu hướng ăn thịt đồng loại gia tăng.²¹

So sánh

Việc tóm tắt lại các tính năng của những cấp hệ thống khác nhau mà chúng ta vừa thảo luận có thể sẽ hữu ích (Bảng 11).

Có thể sẽ cần nghiên cứu thêm để xác định kiểu hệ thống nào là an toàn nhất. Câu trả lời có thể còn tùy thuộc vào các điều kiện triển khai AI. Về mặt an toàn, AI tiên tri rõ ràng là hấp dẫn vì sẽ cho phép áp dụng cả phương pháp kiểm soát năng lực và lựa chọn động lực. Vì thế, cấp AI này xem ra vượt trội so với AI toàn năng, loại chỉ cho phép sử dụng các phương pháp chọn lựa động lực (ngoại trừ những kịch bản mà thế giới được coi là có những siêu trí tuệ mạnh mẽ khác, khi có thể ứng dụng các biện pháp hội nhập xã hội hoặc giam hãm vị nhân). Tuy nhiên, một AI tiên tri có thể mang lại rất nhiều quyền lực cho người vận hành (có thể bị tha hóa hoặc dùng quyền lực đó một cách thiếu suy nghĩ), trong khi một AI toàn năng lại có thể đưa ra một số biện pháp bảo vệ trước những nguy hiểm này. Vì thế, đánh giá mức độ an toàn là một việc không dễ dàng.

AI thần đèn có thể được xem như sự kết hợp hài hòa giữa AI tiên tri và AI toàn năng, nhưng không hẳn tốt. Rất có thể nó sẽ thừa hưởng những yếu điểm của cả hai. Trong khi đó, sự an toàn hiển nhiên của một AI công cụ có thể chỉ là ảo tưởng. Để các công cụ đủ linh hoạt để thay thế cho các tác tử siêu trí tuệ, chúng có thể cần phải triển khai các quá trình tìm kiếm nội bộ và lập kế hoạch cực kỳ mạnh. Những hành vi giống tác tử có thể phát sinh như một hệ quả ngoài dự kiến. Trong trường hợp này, sẽ tốt hơn khi thiết kế hệ thống thành tác tử ngay từ đầu để các lập trình viên có thể thấy được dễ dàng hơn tiêu chí nào cuối cùng sẽ xác định đầu ra của hệ thống.

Bảng 11 Những tính chất của các cấp hệ thống khác nhau

Tiên tri	Một hệ thống Hỏi – Đáp Các biến thể: tiên tri lĩnh vực (ví dụ toán học); tiên tri kết quả giới hạn (ví dụ chỉ có các câu trả lời có/không/không quyết, hoặc các xác suất); tiên tri từ chối trả lời câu hỏi nếu chúng dự báo rằng hệ quả của việc trả lời có thể sẽ thỏa mãn “tiêu chí thảm họa” định trước; nhiều tiên tri để đánh giá ngang hàng.	• Áp dụng đầy đủ các phương pháp đóng hộp • Áp dụng đầy đủ phương pháp thuần hóa • Giảm nhu cầu hiểu dự định và quan tâm của con người cho AI (so với thần đèn và AI toàn năng) • Sử dụng các câu trả lời có/không có thể xóa bỏ nhu cầu về một tiêu chí của “sự hữu ích” hoặc “có thông tin” cho các câu trả lời. • Nguồn gốc của quyền lực vĩ đại (có thể cho người vận hành một lợi thế chiến lược quyết định) • Ít được bảo vệ chống lại việc sử dụng ngu ngốc của người vận hành • Các tiên tri không đáng tin cậy có thể được sử dụng để đưa ra những câu trả lời khó tìm nhưng dễ xác minh • Có thể yếu ớt xác minh các câu trả lời bằng cách sử dụng nhiều tiên tri
Thần đèn	Một hệ thống thực thi mệnh lệnh Các biến thể: Các loại thần đèn dùng những “khoảng cách ngoại suy” khác nhau hoặc tuân thủ tính thần thay vì câu chữ của mệnh lệnh; thần đèn trong từng lĩnh vực; thần đèn từ chối tuân lệnh nếu chúng dự đoán hệ quả của việc tuân lệnh có thể thỏa mãn các “tiêu chí thảm họa” nào đó đã được định trước.	• Áp dụng một phần được các phương pháp đóng hộp (đối với các thần đèn bị hạn chế về mặt không gian) • Áp dụng một phần được phương pháp thuần hóa • Thần đèn có thể cho ta thấy được những khía cạnh quan trọng của các kết quả kỳ vọng • Thần đèn có thể triển khai thay đổi theo các giai đoạn, tạo cơ hội rà soát tại từng giai đoạn • Là nguồn của quyền năng vĩ đại (có thể cho người vận hành một ưu thế chiến lược quyết định) • Ít được bảo vệ chống lại việc sử dụng thiếu suy xét của người vận hành • AI cần thấu hiểu hơn những lợi ích và các dự định của con người (so với các tiên tri)
Toàn năng	Một hệ thống được thiết kế để vận hành độc lập, phóng khoáng Các biến thể: Nhiều hệ thống động lực có thể; khả năng sử dụng xem trước và “phê chuẩn của người bảo trợ” (sẽ được thảo luận ở Chương 13)	• Không áp dụng được các phương pháp đóng hộp • Hầu hết các phương pháp kiểm soát năng lực khác cũng không áp dụng được (có thể ngoại trừ hội nhập xã hội hoặc giam hãm vị nhân) • Gần như không thuần hóa được • AI rất cần hiểu được những lợi ích và dự kiến thực sự của con người • Cần làm đúng ngay từ đầu (mặc dù điều này đúng với tất cả các cấp, có thể ở mức độ thấp hơn) • Có tiềm năng trở thành một quyền năng vĩ đại cho người tài trợ; bao gồm ưu thế chiến lược quyết định • Khi được kích hoạt, có thể bị người vận hành chiếm quyền và có thể được thiết kế với mức độ bảo vệ nhất định chống lại việc sử dụng thiếu suy xét của người vận hành • Có thể sử dụng để triển khai các kết quả “tầm màn vô minh” (xem Chương 13) • Các phương pháp đóng hộp có thể áp dụng được, tùy vào cách triển khai • Các quy trình tìm kiếm mạnh mẽ có thể liên quan trong khi phát triển và vận hành một siêu trí tuệ máy • Tìm kiếm mạnh mẽ để tìm ra một giải pháp đáp ứng được một vài tiêu chí chính thức có thể tạo ra những giải pháp đáp ứng được tiêu chí theo cách nguy hiểm và không mong muốn • Tìm kiếm mạnh mẽ có thể liên quan đến các quá trình tìm kiếm thứ cấp, nội bộ và quy hoạch có thể tìm ra những phương pháp nguy hiểm để thực thi các quá trình tìm kiếm sơ cấp.

Chương 11

Những kịch bản đa cực

Chúng ta đã thấy (đặc biệt trong Chương 8) một kết quả đơn cực đáng sợ đến mức nào, một kết quả mà trong đó, một siêu trí tuệ duy nhất có được ưu thế chiến lược quyết định và sử dụng nó để thiết lập thể đơn nhất. Trong chương này, chúng ta sẽ xem xét điều xảy ra trong một kết quả đa cực, một xã hội hậu chuyển đổi với nhiều đại diện siêu trí tuệ cạnh tranh nhau. Lợi ích của chúng ta trong loại kịch bản này có hai mặt. Thứ nhất, như đã nói qua trong Chương 9, hội nhập xã hội có thể được xem như cung cấp một giải pháp cho vấn đề kiểm soát. Chúng ta đã ghi nhận một số hạn chế của cách tiếp cận này, và chương này sẽ vẽ một bức tranh đầy đủ hơn. Thứ hai, cho dù không ai có dự định tạo ra một điều kiện đa cực để giải quyết vấn đề kiểm soát thì kết quả này vẫn có thể xảy ra. Vậy kết quả đó sẽ ra sao? Xã hội cạnh tranh sinh ra từ đó không nhất thiết phải là xã hội hấp dẫn hay trường tồn.

Trong các kịch bản xuất hiện thể đơn nhất, điều xảy ra sau chuyển đổi gần như phụ thuộc hoàn toàn vào các giá trị của chính thể đơn nhất đó. Kết quả vì thế có thể rất tốt hoặc rất xấu, tùy vào những giá trị này. Tiếp đó, bản thân việc những giá trị này là gì lại phụ thuộc vào việc vấn đề kiểm soát có được giải quyết hay không,

và vào mục tiêu của dự án đã tạo ra thể đơn nhất (theo mức độ mà vấn đề kiểm soát được giải quyết).

Vì vậy, một người quan tâm đến kết quả của các kịch bản về thể đơn nhất thực sự chỉ có ba nguồn thông tin: thông tin về những thứ không thể bị ảnh hưởng bởi hành động của thể đơn nhất (như các định luật vật lý); thông tin về những giá trị hội tụ công cụ; và thông tin cho phép người đó dự báo hoặc giả định xem những giá trị cuối cùng mà thể đơn nhất sẽ có là gì.

Trong các kịch bản về thể đơn nhất, ta còn có một tập hợp những trở ngại khác liên quan đến cách tương tác của các tác tử. Các động lực xã hội xuất hiện từ những tương tác này có thể được nghiên cứu bằng những kỹ thuật từ lý thuyết trò chơi, kinh tế học và thuyết tiến hóa. Các phần tử của khoa học chính trị và xã hội học cũng phù hợp chừng nào còn có thể được tinh lọc và chiết xuất từ một số đặc tính ngẫu nhiên hơn của trải nghiệm con người. Mặc dù kỳ vọng rằng những cản trở này cho chúng ta một bức tranh chính xác của thế giới hậu chuyển đổi là một việc phi thực tế, nhưng chúng vẫn có thể giúp chúng ta xác định một số khả năng cao và chống lại các giả định không có căn cứ.

Chúng ta sẽ bắt đầu bằng cách khám phá một kịch bản kinh tế với đặc trưng là mức độ quản lý thấp, các quyền sở hữu được bảo vệ mạnh mẽ và các dạng thức trí óc kỹ thuật số giá cả phải chăng liên tục xuất hiện.¹ Loại mô hình này gắn với tên tuổi của nhà kinh tế người Mỹ Robin Hanson, người đã thực hiện các công trình tiên phong. Trong chương này, chúng ta sẽ xem xét một số vấn đề về tiến hóa và nghiên cứu các viễn cảnh của một thế giới đa cực hậu chuyển đổi mà sau đó được thống nhất thành một thể đơn nhất.

Chuyện ngựa và người

Trí tuệ máy tổng thể có thể thay thế cho trí tuệ con người. Các trí óc kỹ thuật số không những có thể làm các công việc trí óc hiện nay do con người thực hiện mà còn có thể thay thế được lao động chân tay một khi được trang bị những cánh tay máy hay một cơ thể robot chất lượng tốt. Giả sử các công nhân máy – thứ có thể tái tạo nhanh chóng – trở nên rẻ hơn và có năng lực cao hơn các công nhân người trong hầu hết các loại hình công việc. Chuyện gì sẽ xảy ra khi đó?

Tiền lương và việc làm

Với nhân công rẻ và có thể sao chép, mức tiền lương trên thị trường sẽ giảm. Lĩnh vực duy nhất con người vẫn còn năng cạnh tranh có thể là lĩnh vực mà ở đó, khách hàng cơ bản yêu thích những công việc do chính con người thực hiện. Ngày nay, những mặt hàng thủ công hay do người dân bản xứ sản xuất đôi khi có giá bán cao hơn. Tương tự, người tiêu dùng tương lai có thể ưa chuộng hàng hóa do con người làm ra hay những vận động viên, nghệ sĩ, người tình và lãnh đạo là con người hơn các đồng nhiệm nhân tạo có chức năng tương tự hoặc trội hơn. Tuy nhiên, vẫn chưa rõ là sự ưa chuộng này sẽ rộng rãi đến mức độ nào. Nếu những lựa chọn thay thế do máy móc làm ra đủ vượt trội, có thể chúng sẽ được đánh giá cao hơn.

Một tham số có thể phù hợp với lựa chọn của người tiêu dùng là đời sống tinh thần của công nhân cung cấp một dịch vụ hay sản phẩm nhất định. Chẳng hạn, khán thính giả khi nghe hòa nhạc có thể muốn biết rằng người nghệ sĩ đang thực sự trải nghiệm âm nhạc và không gian nghệ thuật một cách có ý thức. Thiếu đi trải nghiệm sâu sắc, các nghệ sĩ có thể chỉ được coi như những máy chơi nhạc cao

cấp, dù vẫn có thể tạo ra sự hiện diện ba chiều biết tương tác tự nhiên với đám đông. Khi đó máy móc có thể được thiết kế để tạo ra cùng một loại trạng thái tinh thần mà con người có thể có khi thực hiện cùng một tác vụ. Ngay cả khi nhân bản hoàn hảo các trải nghiệm chủ quan, một số người vẫn có thể ưa chuộng các sản phẩm hữu cơ hơn. Sự ưa chuộng này có thể có nguồn gốc ý thức hệ hay tôn giáo. Giống như nhiều người Hồi giáo hay người Do Thái tránh ăn thực phẩm được nấu theo các phương pháp mà họ coi là haram hay treif (cấm kỵ hoặc không phù hợp), trong tương lai sẽ có những nhóm người tránh những sản phẩm mà trong quá trình sản xuất sử dụng trái phép trí tuệ máy.

Hệ quả của nó là gì? Tùy theo mức độ lao động máy móc rẻ tiền có thể thay thế sức người, việc làm của con người có thể biến mất. Những nỗi sợ hãi về tự động hóa và mất việc làm tất nhiên không phải là mới. Quan ngại về thất nghiệp do công nghệ đã từng định kỳ xuất hiện, ít nhất từ Cách mạng Công nghiệp; và thực sự là khá nhiều nghề đã biến mất theo cách những người thợ dệt và thợ thủ công ngành dệt từng đoàn kết lại trong những năm đầu thế kỷ 19 dưới ngọn cờ của “Tướng Ludd” để chống lại việc sử dụng máy dệt cơ giới hóa. Mặc dù vậy, tuy máy móc và công nghệ đã thay thế nhiều loại hình lao động cụ thể của con người nhưng công nghệ vật lý về tổng thể vẫn chỉ là công cụ hỗ trợ lao động. Tiền lương trung bình của con người trên toàn thế giới luôn tăng trưởng về dài hạn, và phần lớn là nhờ những công cụ hỗ trợ này. Nhưng thứ khởi đầu như một công cụ hỗ trợ lao động lại có thể trở thành nguồn lao động thay thế trong tương lai. Ngựa ban đầu được hỗ trợ bằng xe ngựa và cày, những thứ làm năng suất tăng đáng kể, rồi lại bị thay thế bởi xe hơi

và máy cày. Những phát minh mới này làm giảm nhu cầu về ngựa, dẫn đến việc sụt giảm số lượng ngựa. Liệu điều tương tự có xảy ra với loài người?

Sự tương đồng với câu chuyện về ngựa có thể tiếp diễn nếu chúng ta hỏi vì sao loài ngựa vẫn tồn tại. Một lý do là vẫn còn một vài thị trường ngách, trong đó loài ngựa có ưu thế chức năng; chẳng hạn như trong ngành cảnh sát. Nhưng lý do chính là con người có những sở thích khác thường đối với dịch vụ liên quan đến loài ngựa, bao gồm cưỡi ngựa giải trí và đua ngựa. Những sự ưa chuộng này có thể so sánh được với giả định của chúng ta rằng một số người trong tương lai sẽ ưa chuộng những hàng hóa và dịch vụ do chính bàn tay con người làm ra. Cho dù chỉ là gợi ý, nhưng so sánh này cũng không chính xác vì không có thay thế chức năng hoàn toàn nào cho ngựa. Nếu như có những thiết bị cơ khí không đắt tiền có thể chạy hay có hình dáng, cảm giác, mùi và hành vi giống hệt như ngựa thật – và thậm chí còn có cùng trải nghiệm có ý thức – thì khi đó nhu cầu đối với ngựa sinh học có thể giảm hơn nữa.

Với một mức giảm phù hợp trong nhu cầu đối với lao động của con người thì tiền lương có thể giảm xuống dưới mức tối thiểu đủ sống. Ảnh hưởng xấu tiềm năng đối với công nhân người không chỉ là cắt giảm tiền lương, giáng chức hay cần đào tạo lại mà cả đói kém và cái chết. Khi ngựa trở nên lạc hậu về mặt sức kéo, nhiều cá thể đã bị bán để làm thức ăn cho chó, phân bón, da và hồ. Những con vật này không có nghề thay thế để tiếp tục tồn tại. Ở Mỹ năm 1915 có 26 triệu con ngựa, nhưng tới đầu những năm 1950 chỉ còn lại hai triệu con.²

Vốn và phúc lợi

Một sự khác biệt giữa người và ngựa là người sở hữu vốn. Ở đây chúng tôi đưa ra một tình huống thực nghiệm đã được đơn giản hóa: tổng tỷ trọng của vốn như một nhân tố trong sản xuất luôn giữ ở mức cố định khoảng 30% trong một thời gian khá dài (mặc dù có những biến động ngắn hạn khá đáng kể),³ nghĩa là 30% tổng thu nhập toàn cầu là tiền trả cho chủ sở hữu vốn còn 70% còn lại là tiền lương trả cho công nhân. Nếu chúng ta phân loại AI là vốn thì cùng với sự ra đời của trí tuệ máy có thể hoàn toàn thay thế công việc của con người, tiền lương sẽ giảm tới mức chi phí cận biên của những máy móc thay thế này. Mức này – theo giả định rằng máy móc rất hiệu quả – có thể thấp hơn nhiều so với thu nhập đủ sống tối thiểu của con người. Phần thu nhập do lao động nhận được khi đó có thể thực sự giảm xuống bằng không. Nhưng điều này hàm ý rằng tỷ lệ phân chia lãi từ vốn có thể lên tới gần 100% cho tổng sản phẩm thế giới. Do GDP của thế giới sẽ tăng vọt sau một sự bùng nổ trí tuệ (vì số lượng khổng lồ máy móc mới thay thế nhân công nhưng cũng bởi những phát triển công nghệ do siêu trí tuệ tạo ra, và sau này là việc chiếm đoạt lượng khổng lồ đất đai mới nhờ thuộc địa hóa không gian), kết quả là tổng thu nhập từ vốn sẽ tăng lên rất cao. Nếu con người tiếp tục là chủ sở hữu của nguồn vốn này, tổng thu thập dân số nhận được sẽ tăng khủng khiếp, mặc dù trong kịch bản này con người không còn nhận được bất kỳ khoản lương bổng nào.

Loài người nói chung có thể vì thế mà trở nên giàu có hơn mọi ước muốn tham lam. Nhưng thu nhập này sẽ được phân chia ra sao? Theo đánh giá ban đầu, thu nhập từ vốn sẽ tỷ lệ thuận với tổng vốn sở hữu. Với hiệu ứng khuếch đại khổng lồ, ngay cả một lượng tài sản nhỏ tiền chuyển đổi cũng sẽ sinh sôi nảy nở thành một lượng tài

sản rất lớn hậu chuyển đổi. Tuy nhiên, trong thế giới hiện đại, rất nhiều người không có tài sản. Đó không chỉ là những cá nhân sống trong nghèo khó mà cả một số người có thu nhập cao hoặc những người có vốn con người cao nhưng có giá trị thuần âm. Ví dụ, ở các nước giàu có như Đan Mạch hay Thụy Điển, có đến 30% dân số sở hữu lượng tài sản âm – thường là những người trẻ, trung lưu với một vài tài sản hữu hình và nợ thẻ tín dụng hoặc các khoản vay sinh viên.⁴ Thậm chí nếu các khoản tiết kiệm có thể cho lãi suất siêu cao thì vẫn cần có hạt giống, nghĩa là một số vốn ban đầu nào đó để quá trình gộp lãi có thể bắt đầu.⁵

Mặc dù vậy, thậm chí cả những cá nhân không có tài sản riêng vào lúc khởi đầu quá trình chuyển đổi cũng trở nên siêu giàu có. Những người này tham gia vào các quỹ lương hưu chẳng hạn, bất kể là tư hay công, có thể có vị thế tốt, với điều kiện các quỹ này ít nhất được đầu tư một phần.⁶ Giới vô sản cũng có thể trở nên giàu có thông qua lòng nhân từ của những người thấy giá trị thuần của mình tăng vọt: vì kích thích tài sản vĩ đại nên thậm chí chỉ một phần rất nhỏ được hiến tặng cũng tương đương với một tổng có giá trị tuyệt đối rất lớn.

Sự giàu có cũng vẫn có thể đến từ lao động, ngay cả cả ở giai đoạn hậu chuyển đổi khi máy móc vượt trội hơn con người về mặt chức năng trong tất cả các lĩnh vực (cũng như rẻ hơn lao động con người, kể cả ở dưới mức tối thiểu). Như đã lưu ý ở trên, điều này có thể xảy ra nếu có các thị trường ngách, trong đó sức lao động của con người được ưa chuộng vì lý do mỹ thuật, ý thức hệ, đạo đức, tôn giáo hoặc các lý do phi thực dụng khác. Trong một kịch bản mà tài sản của các chủ sở hữu vốn tăng lên đáng kể, nhu cầu đối với lao

động loại này cũng có thể tăng lên tương ứng. Những nghìn tỷ hay vạn tỷ phú mới xuất hiện có thể chi trả những khoản phí lớn để có được một số món hàng hay dịch vụ được cung cấp bởi lực lượng lao động hữu cơ “giao dịch công bằng”. Câu chuyện về ngựa một lần nữa lại cho ta ví dụ tương đương. Sau khi giảm xuống còn hai triệu cá thể trong những năm đầu 1950, số lượng ngựa ở Mỹ đã khôi phục đều đặn: điều tra gần đây cho giá trị gần 10 triệu đầu ngựa.⁷ Sự tăng này không phải do các nhu cầu chức năng mới đối với ngựa trong ngành nông nghiệp hay giao thông mà chính sự tăng trưởng kinh tế đã cho phép nhiều người Mỹ có thể hưởng thụ các môn thể thao giải trí liên quan đến ngựa.

Một sự khác biệt có liên quan khác giữa người và ngựa ngoài việc sở hữu vốn là con người có khả năng huy động chính trị. Một chính phủ do con người điều hành có thể sử dụng quyền thu thuế của nhà nước để phân phối lại lợi nhuận cá nhân, hoặc tạo ra doanh thu bằng cách bán tài sản có giá trị cao thuộc sở hữu nhà nước, ví dụ như đất công, và sử dụng tiền thu được để phục vụ người dân. Và một lần nữa, cũng nhờ sự tăng trưởng kinh tế bùng nổ trong và ngay sau chuyển đổi, sẽ có một lượng tài sản khổng lồ trôi nổi xung quanh, giúp dễ dàng làm đầy bát cơm cho tất cả những công dân thất nghiệp. Thậm chí việc một quốc gia duy nhất cung cấp cho mỗi người trên khắp thế giới một khoản trợ cấp sinh hoạt hào phóng mà chi phí tương ứng không cao hơn những gì nhiều quốc gia đang chi cho viện trợ nước ngoài là hoàn toàn khả thi.⁸

Nguyên tắc Malthus từ quan điểm lịch sử

Cho tới nay chúng ta giả định dân số cố định. Đó có thể là một giả định hợp lý trong một khoảng thời gian ngắn, vì sinh học giới hạn tỷ

lệ tái sinh sản. Tuy nhiên trong những khung thời gian dài hơn, giả định này không phải lúc nào cũng hợp lý.

Dân số loài người đã tăng 2.000 lần trong 9.000 năm qua.⁹ Sự tăng trưởng này có thể còn nhanh hơn nhiều trừ việc trong suốt thời kỳ lịch sử và tiền sử, dân số loài người luôn tăng tới giới hạn của nền kinh tế thế giới. Điều kiện sống gần đúng với lý thuyết Malthus, trong đó hầu hết mọi người nhận được thu nhập dưới mức tồn tại chỉ đủ cho họ sống sót và nuôi được trung bình hai đứa trẻ tới tuổi trưởng thành.¹⁰ Thi thoảng có những đợt “thanh lọc” tạm thời hoặc cục bộ: bệnh dịch, biến động khí hậu hoặc chiến tranh tạm thời loại bỏ một phần dân số và giải phóng đất đai, cho phép những người còn sống cải thiện chế độ dinh dưỡng – và nuôi dạy thêm nhiều con cái hơn, cho tới khi trật tự được khôi phục và điều kiện Malthus tái lập. Ngoài ra, do bất công xã hội, một tầng lớp tinh hoa mỏng manh có thể được hưởng đều đặn khoản thu nhập cao hơn mức đủ sống (với cái giá phải trả là làm giảm phần nào tổng lượng dân số có thể duy trì sự sống). Chúng ta có một ý nghĩ buồn và lạc điệu: trong điều kiện Malthus này, với tình trạng sinh sống bình thường trong hầu hết thời gian chúng ta sống trên Trái đất, thì hạn hán, bệnh dịch, tàn sát và bất bình đẳng – theo đánh giá thông thường là những kẻ tử thù của lợi ích con người – lại có thể là những nhà nhân văn vĩ đại nhất: chỉ có chúng mới đôi khi cho phép có được điều kiện sống trung bình hơi tốt hơn một chút so với mức sống bên bờ vực diệt vong.

Đặt lên những biến động địa phương, lịch sử cho thấy một mẫu hình vi mô của một quá trình tăng trưởng kinh tế trước chậm sau nhanh, được tiếp sức bởi sự tích tụ các sáng tạo công nghệ. Nền kinh tế thế giới tăng trưởng kéo theo sự tăng trưởng dân số toàn cầu

tương xứng. (Chính xác hơn, chính dân số lớn xem ra lại làm tăng đáng kể tốc độ tăng trưởng, có thể chủ yếu bằng cách gia tăng trí tuệ tập thể của loài người.¹¹) Tuy nhiên, chỉ từ sau Cách mạng Công nghiệp, quá trình tăng trưởng kinh tế mới trở nên nhanh chóng, đến mức tăng trưởng dân số không còn theo kịp. Thu nhập trung bình do đó bắt đầu tăng, trước hết ở các nước công nghiệp hóa sớm của Tây Âu, rồi tới hầu hết mọi nơi trên thế giới. Thậm chí ở những nước nghèo nhất thế giới ngày nay, thu nhập trung bình cũng cao hơn đáng kể so với mức đủ sống, được biểu hiện qua việc dân số ở của các nước này đang tăng.

Những quốc gia nghèo nhất ngày nay có tăng trưởng dân số nhanh nhất, vì họ vẫn chưa hoàn thành “quá trình chuyển đổi dân số” sang chế độ sinh sản mức thấp như đã diễn ra ở những xã hội phát triển hơn. Các nhà dân số học dự báo rằng dân số thế giới sẽ tăng lên khoảng 9 tỷ người vào giữa thế kỷ này, và sau đó có thể chững lại hoặc giảm xuống do các nước nghèo hơn đã cùng đi vào chế độ sinh sản mức thấp như các nước phát triển.¹² Nhiều nước giàu đã có tỷ lệ sinh thấp hơn mức thay thế (một số trường hợp thấp hơn rất nhiều).¹³

Nhưng vẫn còn những lý do khác để kỳ vọng về việc có thể trở lại điều kiện bình thường về mặt lịch sử và sinh thái của dân số thế giới để có thể bùng nổ một lần nữa và vượt qua những giới hạn hiện tại, nếu chúng ta nhìn xa hơn và giả định rằng tình trạng công nghệ không thay đổi và thịnh vượng liên tục. Nếu điều này có vẻ phản trực giác khi nhìn từ quan điểm của mối quan hệ ngược giữa tài sản và khả năng sinh sản mà chúng ta đang quan sát thấy trên quy mô toàn cầu, thì chúng ta cần phải tự nhắc nhở mình rằng kỷ nguyên hiện đại

này là một lát cắt mỏng và rất lầm lạc của lịch sử. Hành vi của loài người vẫn chưa được thích ứng với những điều kiện hiện đại. Không những chúng ta thất bại trong việc tận dụng những biện pháp khá hiển nhiên để tăng khả năng thích ứng bao trùm của mình (ví dụ như hiến tặng trứng hay tinh trùng), mà còn tích cực hủy hoại khả năng sinh sản của mình bằng cách kiểm soát sinh sản. Trong môi trường của sự thích ứng tiến hóa, một sự ham muốn tình dục lành mạnh có thể là đủ để khiến cho một cá nhân hành động theo hướng cực đại hóa tiềm năng sinh sản của mình; tuy nhiên trong môi trường hiện đại, còn có một ưu điểm chọn lọc rất lớn để có được một khát khao trực tiếp hơn của việc trở thành cha mẹ sinh học của một số lượng con nhiều nhất có thể. Khát khao này hiện đang được lựa chọn, cũng như các tố chất khác làm tăng khuynh hướng tái tạo của chúng ta. Tuy nhiên, thích ứng văn hóa có thể qua mặt tiến hóa sinh học. Một số cộng đồng, chẳng hạn những cộng đồng Hutterites hoặc những người theo phong trào truyền giáo/ giáo phái Quiverfull, có văn hóa sinh đẻ khuyến khích các gia đình lớn, và sau đó họ đã mở rộng nhanh chóng.

Tăng trưởng dân số và đầu tư

Nếu chúng ta hình dung những điều kiện kinh tế xã hội hiện tại bị đóng băng trong hình thức hiện nay của nó, thì tương lai sẽ bị thống trị bởi các nhóm văn hóa hay dân tộc duy trì được mức độ sinh sản cao. Nếu hầu hết người dân có những ưu tiên để thích ứng tối đa trong môi trường hiện đại thì dân số có thể dễ dàng tăng gấp đôi sau mỗi thế hệ. Nếu thiếu đi các chính sách kiểm soát dân số – thứ có thể liên tục trở nên cứng rắn và hiệu quả hơn trước tinh thần chống đối chính sách ngày càng mạnh mẽ – thì dân số thế giới có thể tiếp tục

tăng trưởng theo cấp số nhân cho tới khi có một trở ngại nào đó, ví dụ như khan hiếm đất đai hoặc cạn kiệt các cơ hội dễ dàng cho những sáng tạo quan trọng: lúc đó, thu nhập trung bình sẽ bắt đầu giảm đến khi rơi xuống mức mà sự nghèo khổ nặng nề sẽ không cho phép hầu hết mọi người nuôi được hai đứa trẻ tới tuổi trưởng thành. Như vậy nguyên tắc Malthus lại tự khẳng định, như một tay chủ nô đáng sợ, và kết thúc cuộc trốn chạy vào miền đất hứa đủ đầy của chúng ta và đưa chúng ta quay lại với xiềng xích, ở đó chúng ta tiếp tục cuộc tranh đấu mệt mỏi để tồn tại.

Viễn cảnh dài hạn hơn này có thể được biến thành một viễn cảnh hiện hữu hơn nữa bởi sự bùng nổ trí tuệ. Vì phần mềm có thể sao chép được, một số lượng các bản giả lập hay AI có thể nhanh chóng tăng gấp đôi – chỉ trong vài phút thay vì hàng thập kỷ hay thế kỷ – nhanh chóng vắt kiệt toàn bộ phần cứng sẵn có.

Tài sản tư nhân có thể phần nào giúp bảo vệ khỏi sự xuất hiện của một điều kiện Malthus phổ quát. Hãy nghiên cứu một mô hình đơn giản, trong đó các bộ tộc (hay những cộng đồng đóng hoặc quốc gia) khởi đầu với lượng tài sản khác nhau và độc lập áp dụng các chính sách khác nhau về sinh sản và đầu tư. Một số bộ tộc coi nhẹ tương lai và nhanh chóng phung phí hết tài sản, sau đó những thành viên nghèo khổ của họ gia nhập hàng ngũ vô sản toàn cầu (hoặc chết nếu không thể tự lao động nuôi thân). Những bộ tộc khác đầu tư một phần nguồn lực của mình nhưng lại áp dụng chính sách sinh sản không giới hạn: những bộ tộc này có thể tăng trưởng ngày càng đông đúc cho tới khi đạt tới điều kiện Malthus nội bộ, khi thành viên của họ nghèo đến nỗi mức tử ngang với mức sinh, và khi đó, tốc độ tăng trưởng dân số của bộ tộc chậm dần xuống tới mức ngang bằng

với tăng trưởng tài nguyên. Trong khi đó, những bộ tộc khác có thể hạn chế mức độ sinh sản của mình dưới mức tăng trưởng vốn của họ: những bộ tộc này có thể dần dần tăng dân số còn thành viên thì trở nên giàu có hơn.

Nếu tài sản được phân phối lại từ các bộ tộc giàu có cho các thành viên của các bộ tộc sinh sản hay tiêu pha nhanh chóng (trẻ em của họ, dù là các bản sao hay hậu duệ thực sự, mặc dù không phải do lỗi của chúng, vẫn bị quăng vào thế giới với số vốn không đủ để tồn tại hay thịnh vượng) thì khi đó họ sẽ tiệm cận hơn với điều kiện Malthus phổ quát. Trong trường hợp giới hạn, thành viên của tất cả các bộ tộc có thể có thu nhập ở mức đủ sống và mọi người đều như nhau về sự nghèo khó.

Nếu tài sản không được phân phối lại, các bộ tộc thông thái có thể nắm giữ một lượng vốn nhất định, và có thể tài sản của họ có thể tăng theo giá trị tuyệt đối. Tuy nhiên, vẫn chưa rõ liệu con người có được tỷ suất hoàn vốn cao như các trí tuệ máy có thể có được từ vốn của chúng hay không, bởi sẽ có thể có sự kết hợp giữa lao động và vốn sao cho một tác tử duy nhất cung cấp được cả hai (ví dụ một nhà kinh doanh hoặc nhà đầu tư vừa có kỹ năng vừa giàu có) đạt được tỷ lệ hoàn vốn riêng cho vốn của mình cao hơn tỷ lệ thị trường mà các tác tử sở hữu nguồn lực tài chính nhưng không sở hữu nguồn lực nhận thức có thể có được. Con người, vốn có kỹ năng không bằng trí tuệ máy, có thể vì thế mà gia tăng vốn chậm chạp hơn – tất nhiên trừ khi vấn đề kiểm soát đã được hoàn toàn giải quyết, và khi đó, tỷ lệ hoàn vốn của con người và máy móc sẽ bằng nhau, vì một người chủ có thể yêu cầu các tác tử máy quản lý tiền tiết kiệm mà không tốn chi phí và và không bị xung đột lợi ích. Nhưng ngược

lại, trong kịch bản này, phần kinh tế do máy móc sở hữu có thể đạt tới 100% một cách bất đối xứng.

Một kịch bản mà trong đó, nền kinh tế do máy móc sở hữu tiếp cận con số 100% một cách bất đối xứng không nhất thiết phải là kịch bản mà phần của loài người bị suy giảm. Nếu kinh tế tăng trưởng với tốc độ thích hợp thì ngay cả một phần đang suy giảm của nó cũng có thể có giá trị tuyệt đối tăng lên. Việc này có vẻ như là một tin tương đối tốt cho loài người: trong một kịch bản đa cực mà quyền sở hữu được bảo vệ – kể cả khi chúng ta hoàn toàn thất bại trong việc giải quyết các vấn đề kiểm soát – tổng tài sản do con người sở hữu có thể tăng. Tất nhiên, hiệu ứng này chưa chú ý đến việc sự tăng trưởng dân số của loài người kéo thu nhập đầu người xuống mức chỉ đủ sống hoặc vấn đề con người tự hủy hoại mình vì họ đã lạm tiêu tương lai.

Trong dài hạn, nền kinh tế có thể bị thống trị bởi những dòng tộc có tỷ lệ tiết kiệm cao nhất – những kẻ keo kiệt sở hữu nửa thành phố nhưng lại ngủ dưới gầm cầu. Chỉ tới khi hết thời gian, khi không còn cơ hội đầu tư nào nữa thì những kẻ keo kiệt thịnh vượng tối đa này mới rút tiền tiết kiệm của mình.¹⁴ Tuy nhiên, nếu bảo vệ quyền sở hữu không được hoàn hảo – ví dụ nếu những máy móc hiệu quả hơn trên mạng thành công, bằng cách này hay cách khác, trong việc chuyển tài sản từ con người sang cho chúng – khi đó các nhà tư bản có thể phải tiêu lạm vào vốn của mình sớm hơn trước khi bị hành động chuyển tài sản (hoặc chi phí thường xuyên để bảo vệ tài sản của họ khỏi hành động đó) làm cho cạn kiệt. Nếu những diễn biến này xảy ra trong thang thời gian số thay vì thời gian sinh học, thì con

người kỹ bằng hà có thể thấy mình trắng tay trước khi kịp thốt lên nửa lời.¹⁵

Cuộc sống trong một nền kinh tế thuật toán

Cuộc đời của những con người sinh học trong trạng thái hậu chuyển đổi Malthus không nhất thiết phải giống với bất cứ trạng thái lịch sử nào của loài người (như người săn bắt hái lượm, nông dân hay nhân viên văn phòng). Thay vào đó, phần lớn con người trong kịch bản này có thể là những người hưu trí rảnh rỗi, gồng gượng qua ngày với khoản tiền lương hưu ít ỏi.¹⁶ Họ có thể rất nghèo, và nhận những khoản thu nhập nhỏ nhoi của mình từ lãi suất tiết kiệm hoặc trợ cấp của chính phủ. Họ có thể sống trong một thế giới với công nghệ cực kỳ tiên tiến, bao gồm không chỉ siêu trí tuệ mà còn thuốc chống lão hóa, thực tại ảo, các công nghệ tăng cường và các loại “thuốc sững” khác nhau: mặc dù vậy những thứ đó vẫn ngoài tầm với. Có thể thay vì sử dụng dược phẩm tăng cường, họ có thể uống thuốc ngăn ngừa sinh trưởng và làm chậm quá trình trao đổi chất để giảm chi phí sinh hoạt (những người tiêu hóa tốt có thể không sống sót được với thu nhập giảm dần). Khi dân số loài người tăng lên và thu nhập trung bình tiếp tục giảm, chúng ta có thể dần suy sụp xuống tới bất cứ cấu trúc tối thiểu nào vẫn còn được phép nhận lương hưu – có thể là những bộ não có nhận thức tối thiểu ngâm trong ống nghiệm, được máy móc cấp dưỡng khí và thức ăn, chậm chạp tiết kiệm cho đủ tiền sinh sản bằng cách để kỹ thuật viên robot phát triển bản sao của chính họ.¹⁷

Có thể căn cơ hơn nữa bằng cách tải lên, vì một chất nền tính toán được tối ưu về mặt vật lý do siêu trí tuệ tiên tiến tạo ra có thể hiệu quả hơn một bộ não sinh học. Tuy nhiên, việc dịch chuyển vào

lãnh địa số có thể bị ngăn cản nếu các bản giả lập bị xem là không phải người hay không phải công dân và không có quyền nhận lương hưu hoặc sở hữu các tài khoản tiết kiệm được miễn thuế. Trong trường hợp này, một ngách dành cho con người sinh học vẫn có thể tiếp tục rộng mở, có thể cùng với một lượng lớn các bản giả lập hay trí tuệ nhân tạo.

Hiện chúng ta đang tập trung vào số phận những người có thể được hỗ trợ bởi các khoản tiết kiệm, trợ cấp hoặc thu nhập tiền lương từ những người khác muốn thuê nhân công. Bây giờ hãy chuyển sang một số thực thể mà cho tới nay chúng ta vẫn phân loại là “vốn”: Những máy móc có thể thuộc sở hữu của con người, được chế tạo và vận hành vì những tác vụ chức năng mà chúng thực hiện và có khả năng thay thế lao động của con người trong rất nhiều loại công việc. Tình huống sẽ như thế nào đối với những “con ngựa” này của nền kinh tế mới?

Nếu những máy móc này là những thiết bị tự động, đơn giản như một động cơ hơi nước hoặc bộ máy đồng hồ, thì không cần phải bình luận gì thêm: sẽ có một lượng lớn loại vốn tư bản này trong một nền kinh tế hậu chuyển đổi, nhưng không ai buồn quan tâm đến việc mọi vật sẽ ra sao với các thiết bị vô tri giác này. Tuy nhiên, nếu máy móc có trí óc nhận thức – nếu chúng được chế tạo sao cho hoạt động của chúng được gắn với nhận biết hiện tượng (hoặc nếu vì một số lý do nào đó chúng được gán cho một vị thế đạo đức) – thì khi đó việc xem xét kết quả toàn thể từ góc nhìn tác động của nó tới các trí óc máy móc này sẽ là quan trọng. Phúc lợi của các trí óc máy móc làm việc này thậm chí còn có thể là khía cạnh quan trọng nhất của kết quả vì chúng có thể có số lượng rất lớn.

Nô lệ tự nguyện và cái chết tự nhiên

Một câu hỏi quan trọng bước đầu là liệu các trí óc máy móc làm việc có bị sở hữu như vốn (nô lệ) hay được thuê như những người lao động nhận lương tự do. Tuy nhiên khi xem xét gần hơn, ta sẽ có nghi ngờ rằng liệu có bất cứ điều gì phụ thuộc vào vấn đề này không. Điều này có hai lý do. Thứ nhất, nếu một công nhân tự do trong một nhà nước Malthus được trả lương ở mức đủ sống, anh ta sẽ không còn thu thập sau khi đã chi vào thức ăn và những nhu yếu phẩm khác. Nếu thay vào đó, người này là một nô lệ, chủ nô sẽ chi trả tiền sinh sống và anh ta cũng vẫn không có thu nhập khả dụng. Trong cả hai trường hợp, người công nhân không nhận được gì khác ngoài nhu yếu phẩm. Thứ hai, giả sử người lao động tự do bằng cách nào đó có khả năng nhận được khoản thu nhập cao hơn mức đủ sống (có thể là nhờ vào những quy định thuận lợi). Anh ta sẽ chi tiêu khoản thặng dư này như thế nào? Các nhà đầu tư sẽ thấy rằng tạo ra những công nhân có thể trở thành “nô lệ tự nguyện” – những người sẵn sàng làm việc vì những khoản tiền lương ở mức đủ sống – là việc làm đem lại lợi nhuận cao nhất. Các nhà đầu tư có thể tạo ra những công nhân như vậy bằng cách sao chép những công nhân tuân thủ. Với lựa chọn thích hợp (và có thể một chút sửa đổi mã), các nhà đầu tư có thể tạo ra những công nhân không chỉ tình nguyện lao động mà còn lựa chọn ủng hộ lại cho chủ mình bất cứ khoản thu nhập thặng dư nào mà họ có thể nhận được. Đưa tiền cho công nhân khi đó sẽ là đường vòng để đưa tiền cho chủ sở hữu hay người sử dụng lao động, ngay cả nếu công nhân là một tác tử tự do với đầy đủ quyền hợp pháp.

Điều này có thể sẽ bị phản đối rằng để thiết kế được một chiếc máy muốn tự nguyện làm bất cứ công việc nào được giao hay ủng hộ tiền lương cho chủ là rất khó. Cụ thể, các bản giả lập có thể được hình dung như những thực thể có những ham muốn gần với con người hơn. Nhưng hãy lưu ý rằng ngay cả nếu vấn đề kiểm soát ban đầu là rất khó khăn thì ở đây chúng ta đang xem xét một điều kiện sau khi chuyển đổi, khi mà các phương pháp lựa chọn động lực được giả định là đã hoàn thiện. Với các bản giả lập, ta có thể đi được khá xa chỉ đơn giản bằng cách chọn lọc từ những cá tính sẵn có của con người; và chúng ta đã mô tả một vài phương pháp lựa chọn động lực. Bài toán kiểm soát có thể được đơn giản hóa bằng cách nào đó bởi giả định hiện tại rằng trí tuệ máy mới gia nhập vào ma trận kinh tế xã hội ổn định đã tràn đầy những tác tử siêu trí tuệ tuân thủ luật pháp khác.

Sau đây, hãy cùng xem xét hoàn cảnh khó khăn của chiếc máy thuộc tầng lớp lao động, bất kể hoạt động như một nô lệ hay một tác tử tự do. Trước hết chúng ta sẽ tập trung vào các bản giả lập, tình huống dễ hình dung nhất.

Để đưa một công nhân sinh học mới “vào đời” tốn đâu đó từ 15 đến 30 năm, tùy thuộc vào trình độ và kinh nghiệm cần thiết. Trong thời gian đó, con người mới này cần được cho ăn, cho ở, chăm sóc và dạy dỗ – với chi phí tốn kém. Ngược lại, tạo ra một bản sao mới của một công nhân số lại dễ dàng như tải một chương trình mới vào bộ nhớ hoạt động. Do đó, sự sống trở nên rẻ rúng. Một doanh nghiệp có thể tùy chỉnh lực lượng lao động của mình để đáp ứng nhu cầu bằng cách sinh ra những bản sao mới và tiêu hủy các bản sao không còn cần đến nữa để giải phóng nguồn lực máy tính. Điều này có thể

dẫn đến tỷ lệ tử rất cao trong các công nhân kỹ thuật số. Nhiều cá thể trong đó chỉ sống được một ngày theo thời gian chủ quan (subjective day).

Còn có những lý do khác với sự biến động trong nhu cầu khiến cho người sử dụng lao động hay chủ của các mô phỏng có thể muốn thường xuyên “giết” hoặc “kết liễu” công nhân của họ.¹⁸ Nếu một trí óc giả lập, như một trí óc sinh học, thì cần các khoảng ngủ hoặc nghỉ ngơi để có thể vận hành, có thể việc xóa bỏ một bản giả lập đã mệt mỏi vào cuối ngày làm việc và thay bằng một trạng thái lưu trữ tươi mới và khỏe khoắn sẽ rẻ tiền hơn. Do quá trình này có thể tạo ra sự mất trí nhớ thụt lùi đối với mọi thứ học được trong ngày hôm đó, các bản giả lập thực hiện những tác vụ cần tới các luồng nhận thức dài phải được bảo vệ chống lại việc xóa bộ nhớ thường xuyên này. Ví dụ, hẳn sẽ khó để viết một cuốn sách nếu mỗi buổi sáng khi ngồi vào bàn, ta không còn ký ức về những gì mình từng làm trước đó. Nhưng các công việc khác có thể được thực hiện một cách thích hợp bởi các tác tử được “tái chế” thường xuyên: một trợ lý bán hàng hay đại diện dịch vụ khách hàng, một khi được đào tạo, có thể chỉ cần nhớ thông tin mới trong 20 phút.

Vì tái chế các bản giả lập có thể cản trở việc hình thành ký ức và kỹ năng, một số bản giả lập có thể được đặt vào một chương trình học tập đặc biệt, ở đó chúng sẽ chạy liên tục cả lúc nghỉ, ngay cả trong những công việc không thực sự cần đến các luồng nhận thức dài. Ví dụ, một số tác tử chăm sóc khách hàng có thể chạy trong nhiều năm trong những môi trường học tập được tối ưu hóa, được huấn luyện viên và những chuyên gia đánh giá hiệu năng hỗ trợ. Những “học viên” giỏi nhất sẽ được dùng làm khuôn mẫu để in ra

hàng triệu bản sao mỗi ngày. Rất nhiều nỗ lực sẽ được đổ vào việc cải thiện hiệu năng của các khuôn mẫu công nhân, vì một sự tăng trưởng dù là nhỏ của năng suất lao động cũng tạo ra giá trị kinh tế khổng lồ khi áp dụng vào hàng triệu bản sao.

Song song với nỗ lực đào tạo các khuôn mẫu công nhân cho công việc cụ thể, nhiều nỗ lực to lớn cũng được thực hiện để cải thiện công nghệ giả lập nền tảng. Tiến bộ ở đây có thể còn có giá trị cao hơn tiến bộ trong các khuôn mẫu công nhân riêng lẻ, vì những cải thiện công nghệ tổng quát có thể được áp dụng cho tất cả các công nhân giả lập (và có thể cho cả các mô phỏng không phải công nhân) thay vì chỉ cho những mô phỏng trong một ngành nghề cụ thể. Những nguồn lực lớn có thể được dành để tìm kiếm những lối tắt tính toán cho phép triển khai hiệu quả hơn những mô phỏng hiện tại, và còn vào việc phát triển các kiến trúc AI dạng thần kinh hoặc hoàn toàn tổng hợp. Nghiên cứu này có thể chủ yếu được thực hiện bởi các mô phỏng chạy trên những phần cứng rất nhanh. Tùy vào giá thành của công suất tính toán, hàng triệu, hàng tỷ hoặc hàng ngàn tỷ mô phỏng của những trí não nghiên cứu sắc sảo nhất của con người (hoặc các phiên bản được mở rộng của họ) có thể làm việc không ngừng nghỉ để mở rộng cương thổ của trí tuệ máy; và một số trong số đó có thể hoạt động nhanh hơn những bộ não sinh học nhiều cấp khuếch đại.¹⁹ Đó là một lý do tốt để nghĩ rằng kỷ nguyên của các bản giả lập tương tự con người có thể ngắn ngủi – một khoảng lặng ngắn trong thời gian thiên văn – và kỷ nguyên này có thể nhanh chóng nhường cho một kỷ nguyên của trí tuệ nhân tạo vượt trội hơn rất nhiều.

Chúng ta đã nhắc đến một số nguyên nhân của việc chủ lao động của các công nhân giả lập có thể thường xuyên phải thanh lọc đội ngũ: những biến động nhu cầu đối với các loại lao động khác nhau, tiết kiệm chi phí của việc không phải mô phỏng thời gian ngủ nghỉ và sự ra đời của những khuôn mẫu mới và cải tiến. Các vấn đề an ninh có thể cho một nguyên nhân khác. Để ngăn công nhân phát triển các kế hoạch và âm mưu lật đổ, các bản giả lập ở một số vị trí nhạy cảm có thể chỉ được chạy trong những khoảng thời gian giới hạn, được thường xuyên trả về trạng thái sẵn sàng được lưu trữ trước đó.²⁰

Những trạng thái sẵn sàng mà các bản giả lập sẽ thiết lập lại này có thể được chuẩn bị và kiểm định kỹ càng. Một bản giả lập “đoàn thò” tiêu biểu có thể tỉnh dậy trong một trạng thái tinh thần được nghỉ ngơi đầy đủ đã được tối ưu hóa cho lòng trung thành và năng suất lao động. “Anh ta” nhớ rằng mình từng đứng đầu lớp sau nhiều năm (chủ thể) đào tạo và lựa chọn gặt gao, sau đó được hưởng thụ một ngày nghỉ phục hồi và một giấc ngủ say, rồi được nghe một bài nói gây cảm hứng và âm nhạc sôi nổi, và bây giờ đã đang sẵn sàng làm việc để cống hiến toàn bộ sức lực cho chủ mình. “Anh ta” không quá bị xáo động bởi những ý nghĩ về cái chết không thể tránh khỏi của mình vào cuối ngày làm việc. Những công nhân giả lập lo lắng về cái chết hoặc các rối loạn tâm thần khác sẽ có năng suất thấp hơn và ít được chọn lựa hơn.²¹

Công việc hiệu quả tối đa liệu có thú vị?

Một biến số quan trọng trong việc đánh giá mức độ mong muốn của một điều kiện giả thuyết như vậy là trạng thái vui sướng của bản giả lập tiêu biểu.²² Một công nhân giả lập tiêu biểu sẽ thấy khổ sở hay

hưởng thụ trải nghiệm lao động vất vả để giải quyết công việc được giao?

Chúng ta cần kháng cự lại cảm dỗ của việc phóng chiếu tình cảm cá nhân lên phiên bản công nhân giả lập tưởng tượng. Vấn đề là liệu bạn có cảm thấy hạnh phúc hay không nếu như phải liên tục làm việc và không còn được ở bên người thân – thành thật mà nói, đó là một số phận thật khủng khiếp.

Sẽ phù hợp hơn tương đối khi ta xem xét trải nghiệm vui sướng điển hình hiện tại của con người trong giờ làm việc. Các nghiên cứu trên toàn cầu phỏng vấn những người được khảo sát về mức độ hạnh phúc thấy rằng hầu hết đều đánh giá mình là “khá hạnh phúc” hoặc “rất hạnh phúc” (điểm trung bình 3,1 trên thang điểm từ 1 đến 4).²³

Những nghiên cứu về hiệu ứng, hỏi những người được khảo sát gần đây đã trải nghiệm được trạng thái có các hiệu ứng tích cực và tiêu cực khác nhau tới mức nào, thường có được kết quả tương tự (có hiệu ứng thuần đạt khoảng 0,52 trên thang điểm từ -1 đến 1). Mức thu nhập trên đầu người của quốc gia có tác động tích cực ở mức khiêm tốn lên cảm nhận sáng khoái chủ quan trung bình.²⁴ Tuy nhiên, sẽ nguy hiểm nếu ngoại suy từ những kết quả này sang trạng thái vui sướng của những công nhân giả lập trong tương lai. Việc này có thể do điều kiện của họ khác nhau: một mặt, họ có thể làm việc vất vả hơn nhiều, mặt khác họ có thể không mắc bệnh, không đau đớn, đói khát, không bốc mùi, ... Mặc dù vậy những xem xét này cơ bản là không trúng mục tiêu. Xem xét quan trọng hơn nhiều ở đây là tông điệu vui sướng có thể được điều chỉnh nhờ những liệu pháp số tương đương với thuốc hay phẫu thuật thần kinh. Điều này có nghĩa

là sẽ sai lầm nếu ta suy ra trạng thái vui sướng của các bản giả lập tương lai từ những điều kiện bên ngoài của cuộc sống của chúng bằng cách tưởng tượng xem chính chúng ta và những cá nhân khác giống chúng ta có thể cảm thấy thế nào trong những hoàn cảnh đó. Trạng thái vui sướng sẽ chỉ là lựa chọn. Trong mô hình hiện tại mà chúng tôi đang xem xét, lựa chọn có thể được thực hiện bởi các chủ sở hữu vốn đang tìm cách cực đại hóa giá trị hoàn vốn đầu tư của họ vào những công nhân giả lập. Hệ quả là câu hỏi về mức độ hạnh phúc của các bản giả lập sẽ quy về câu hỏi rằng trạng thái vui sướng nào mang lại năng suất cao nhất (trong những công việc khác nhau mà các bản giả lập có thể được sử dụng để thực hiện).

Ở đây, một lần nữa, ta có thể tìm cách rút ra một kết luận từ những quan sát về hạnh phúc của con người. Nếu hầu hết mọi lúc, mọi nơi và mọi ngành nghề, con người đều thường hạnh phúc ít nhất ở mức khiêm tốn, thì điều đó sẽ tạo ra một vài giả định có cùng hàm ý trong một kịch bản hậu chuyển đổi như kịch bản chúng ta đang nghiên cứu. Nói rõ hơn, luận điểm trong trường hợp này không phải là trí óc con người có thiên hướng tìm kiếm hạnh phúc nên họ có thể tìm kiếm sự thỏa mãn trong những điều kiện mới mẻ đó; nói đúng hơn, một mức độ hạnh phúc trung bình nhất định nào đó đã được minh chứng là thích ứng với các trí óc con người trong quá khứ thì có thể mức độ hạnh phúc tương tự cũng sẽ minh chứng là tương thích với những trí óc giống con người trong tương lai. Mặc dù vậy, phát biểu này cũng tiết lộ yếu điểm của kết luận: các trạng thái tinh thần tương thích với người tiền sử săn bắn hái lượm lang thang trên đồng cỏ châu Phi không nhất thiết tương thích với các bản giả lập đã sửa đổi sống trong các thực tại ảo thời hậu chuyển đổi. Chúng ta

chắc chắn có thể hy vọng rằng những công nhân giả lập trong tương lai có thể hạnh phúc bằng hoặc hơn những công nhân tiêu biểu trong lịch sử loài người; nhưng chúng ta vẫn chưa thấy bất cứ nguyên nhân nào đủ hấp dẫn để có thể đưa ra giả định như vậy (trong kịch bản tự do kinh tế đa cực hiện đang được nghiên cứu).

Hãy xem xét khả năng của việc nguyên nhân khiến cho hạnh phúc phổ biến trong loài người (tới bất cứ mức độ nào) là tâm trạng vui vẻ hoạt động như một chức năng báo hiệu trong môi trường của sự thích nghi tiến hóa. Tạo ấn tượng cho những thành viên khác của nhóm xã hội về việc đang ở trong trạng thái thăng hoa – khỏe mạnh, có địa vị tốt so với những người ngang hàng và tin tưởng chắc chắn vào vận may kéo dài – có thể làm tăng sự nổi tiếng của một cá nhân. Một thiên kiến nghiêng về tươi tỉnh sẽ được lựa chọn, kết quả là hóa học thần kinh của con người giờ đây nghiêng về tác động tích cực thay vì thứ có thể mang lại hiệu quả cao nhất theo những tiêu chí vật chất đơn giản hơn. Nếu đúng như vậy thì tương lai của niềm vui cuộc sống (*joie de vivre*) có thể phụ thuộc vào việc tâm trạng tươi tỉnh có khả năng duy trì không đổi chức năng báo hiệu xã hội của nó trong thế giới hậu chuyển đổi: một vấn đề mà chúng ta sẽ quay trở lại ngay.

Nếu những tâm hồn mừng vui tỏa ra nhiều năng lượng hơn những tâm hồn buồn thảm thì sao? Có lẽ những tâm hồn vui vẻ dễ sáng tạo và hưởng thụ hơn – những hành vi mà những người chủ lao động trong tương lai không thích nhất ở hầu hết công nhân của họ. Có lẽ một sự quyết tâm âm thầm và bền chồn vào việc tiếp tục công việc không phạm sai lầm sẽ là thái độ tăng năng suất tối đa trong hầu hết các loại công việc. Chúng ta không tuyên bố ở đây rằng

mọi thứ không phải như vậy, mà rằng mình không biết nó có như vậy hay không. Mặc dù vậy, chúng ta phải nghiên cứu xem mọi thứ sẽ tệ đến đâu nếu một số giả thuyết bi quan về tương lai của nhà nước Malthus lại trở thành sự thực: không chỉ vì chi phí cơ hội của việc không tạo ra được điều gì đó tốt hơn – chi phí này có thể là khổng lồ - mà còn bởi vì chính bản thân nhà nước cũng có thể xấu xa, có thể còn tệ hại hơn nhiều so với nhà nước Malthus ban đầu.

Chúng ta ít khi nỗ lực hết sức. Khi làm như vậy, có lúc ta cảm thấy đau đớn. Hãy tưởng tượng chúng ta chạy trên máy chạy đặt ở độ dốc lớn – tim đập mạnh, các bắp thịt đau đớn, phổi hỗn hển. Một cái liếc nhìn đồng hồ: lần giải lao tiếp theo, cũng chính là lúc ta chết, sẽ đến sau 49 năm, 3 tháng, 20 ngày, 4 giờ, 56 phút và 12 giây. Bạn muốn mình chưa từng được sinh ra. Một lần nữa tuyên bố ở đây không phải là nó sẽ như thế nào, mà là chúng không biết rằng nó sẽ không phải như thế. Một người có thể chắc chắn đưa ra một luận điểm lạc quan hơn. Chẳng hạn, không có lý do rõ ràng nào để các giả lập phải chịu đựng những chấn thương và bệnh tật: việc loại bỏ những khổ sở vật chất có thể là một cải thiện lớn so với tình trạng hiện tại. Hơn nữa, vì thứ mà từ đó thực tại ảo được làm ra có thể khá rẻ tiền, nên các mô phỏng có thể làm việc trong cảnh xa hoa – những cung điện trên đỉnh núi, một mái hiên trong khu rừng đậm chồi nảy lộc mùa xuân, hoặc trên những bãi biển bên bờ vịnh xanh ngấn ngát – với cường độ ánh sáng, nhiệt độ, phong cảnh và trang trí vừa đủ; không bị ảnh hưởng bởi khói bụi, tiếng ồn, rác rưởi và lũ côn trùng vo ve; mặc đồ thoải mái, cảm giác sạch sẽ và tập trung và được ăn uống đầy đủ. Quan trọng hơn nữa, nếu – điều hoàn toàn có thể – trạng thái tâm thần tối ưu của con người khi làm tốt hầu hết các công

việc là một “sự nỗ lực sung sướng”, thì khi đó kỷ nguyên của nền kinh tế mô phỏng có thể mang tính kỷ sinh khá cao.

Trong mọi trường hợp, việc sắp xếp vật chất sao cho có ai hoặc cái gì đó có thể can thiệp để khôi phục trật tự nếu quỹ đạo mặc định chệch hướng tới miền khổ ải sẽ có giá trị vô cùng to lớn. Cũng nên có thêm một lối thoát hiểm nào đó cho phép giải cứu bằng cái chết hoặc lãng quên nếu chất lượng cuộc sống vĩnh viễn rớt xuống dưới mức sống không bằng chết.

Những kẻ thuê ngoài vô thức

Xa hơn nữa, khi kỷ nguyên giả lập nhường chỗ cho một kỷ nguyên trí tuệ nhân tạo (hoặc nếu trực tiếp có được trí tuệ máy qua AI, không cần qua giai đoạn giả lập hoàn chỉnh não bộ) thì đau đớn và khoái lạc có thể có khả năng hoàn toàn biến mất trong một kết quả đa cực vì một cơ chế tưởng thưởng khoái lạc có thể không phải là hệ thống động lực hiệu quả nhất cho một tác tử nhân tạo phức tạp (không phải chịu gánh nặng từ phần não bộ động vật như con người). Có thể một hệ thống động lực tiên tiến hơn phải được dựa trên một phép biểu diễn tường minh của một hàm thỏa dụng hay một kiến trúc khác không có các phép biểu diễn dạng hàm tương ứng với đau đớn và khoái lạc.

Một kết quả đa cực có liên quan nhưng cấp tiến hơn hơn – kết quả có thể loại bỏ gần như toàn bộ giá trị khởi tương lai – là khi toàn bộ giới vô sản đều không có ý thức. Khả năng này rõ ràng nhất đối với AI, thứ có thể có cấu trúc rất khác với trí tuệ con người. Nhưng mặc dù trí tuệ máy ban đầu ra đời nhờ giả lập hoàn chỉnh não bộ, tạo ra các trí óc số có ý thức, những lực lượng cạnh tranh được giải phóng trong một nền kinh tế hậu chuyển đổi có thể dễ dàng dẫn đến

việc xuất hiện của những dạng thức trí tuệ máy khác dần so với hệ thần kinh, lý do vì

AI tổng hợp được chế tạo mới hoặc vì các mô phỏng, có thể bằng con đường liên tục điều chỉnh và cải thiện, ngày càng khác với dạng thức con người ban đầu của chúng. Hãy nghiên cứu một kịch bản mà sau khi công nghệ giả lập đã được phát triển, thì những tiến bộ liên tục trong khoa học thần kinh và khoa học máy tính (có được bởi sự hiện diện của các trí não số, vừa là nghiên cứu viên, vừa là đối tượng nghiên cứu) khiến cho có thể tách biệt những module nhận thức riêng biệt trong một mô phỏng rồi nối chúng với những module được tách ra từ những bản giả lập khác. Có thể cần một khoảng thời gian dành cho đào tạo và điều chỉnh trước khi các module khác nhau có thể cộng tác hiệu quả; nhưng các module tuân thủ tiêu chuẩn chung có thể tương thích nhanh hơn với các module tiêu chuẩn khác. Điều đó khiến cho các module được tiêu chuẩn hóa có năng suất tốt hơn và tạo ra áp lực chuẩn hóa nhiều hơn.

Các bản giả lập có thể bắt đầu “outsource” (thuê ngoài) các chức năng của mình ngày càng nhiều hơn. Tại sao phải học giải thuật khi ta có thể gửi bài toán suy luận số học của mình tới Công ty Gauss-Modules? Vì sao phải mất công diễn giải trong khi ta có thể thuê công ty Coleridge Conversations biến những ý nghĩ của mình thành ngôn từ? Vì sao phải ra các quyết định về cuộc sống cá nhân khi có những module thực hành đã được chứng nhận có thể sao quét hệ thống mục tiêu của bạn và quản lý nguồn lực của bạn để đạt được tốt hơn các mục tiêu so với bạn tự mình thực hiện? Một vài bản giả lập có thể chuộng việc giữ lại hầu hết chức năng của mình và tự thực thi các tác vụ có thể được những bản giả lập khác thực hiện tốt hơn. Các

bản giả lập này có thể giống như những người nghiệp dư thích tự trồng rau ăn hay tự đan áo len cho mình. Những mô phỏng nghiệp dư này có thể kém hiệu quả hơn; và nếu có một dòng chảy tài nguyên thuần từ những người tham gia kém hiệu quả trong nền kinh tế sang những người có hiệu quả cao hơn thì cuối cùng những tay nghiệp dư sẽ thất bại.

Những “viên canh” của thứ trí thức rời rạc giống con người vì thế mà tan vào nôi “xúp thuật toán”.

Hiệu năng tối ưu có thể đạt được bằng cách nhóm các năng lực thành những tổ hợp gần giống kiến trúc nhận thức của trí não con người. Ví dụ, có khả năng một module toán học cần được “may đo” theo một module ngôn ngữ, và hai module này phải được tùy chỉnh theo module quản lý để cả ba làm việc được với nhau. Việc “thuê ngoài” nhận thức khi đó có thể hoàn toàn không thực hiện được. Nhưng khi không có nguyên nhân hợp lý nào để tin rằng mọi thứ sẽ diễn ra như vậy, chúng ta phải chấp nhận khả năng các kiến trúc nhận thức tương tự con người chỉ tối ưu trong khuôn khổ các hạn chế về đặc điểm thần kinh học của con người (hoặc hoàn toàn không tối ưu). Khi có thể xây dựng những kiến trúc không thể triển khai tốt trên các mạng thần kinh sinh học, không gian thiết kế mới mở ra; và những thiết kế tối ưu trong không gian mở rộng này không nhất thiết giống với các loại hình tâm tính quen thuộc. Các tổ chức nhận thức tương tự con người có thể không có chỗ cho mình trong một nền kinh tế hoặc hệ sinh thái cạnh tranh hậu chuyển đổi.²⁵

Có thể có các góc ngách cho các tổ hợp đơn giản hơn (như các module riêng biệt), phức tạp hơn (như các cụm module khổng lồ), hoặc phức tạp tương đương trí não con người nhưng có các kiến

trúc hoàn toàn khác. Liệu những tổ hợp này có giá trị nội tại sẵn có nào hay không? Chúng ta có nên chào đón một thế giới mà trong đó, những tổ hợp khác lạ này thay thế các tổ hợp con người?

Đáp án có thể tùy thuộc vào bản chất đặc thù của những tổ hợp xa lạ này. Thế giới hiện tại có nhiều tầng tổ chức. Một số thực thể có độ phức tạp cao như các công ty đa quốc gia hay nhà nước, có thành viên là con người; mặc dù vậy chúng ta thường chỉ gán giá trị công cụ cho những tổ hợp có độ phức tạp cao này. Các công ty và nhà nước (thường được giả định rằng) không có ý thức cao hơn so với những cá nhân cấu thành: họ không cảm nhận được nỗi đau hay khoái lạc hay trải nghiệm được bất cứ qualia nào. Chúng ta đánh giá các thể chế này theo mức độ phục vụ nhu cầu con người của chúng, và sẽ “giết” chúng không thương tiếc khi chúng không còn làm được như vậy. Ngoài ra còn có những thực thể cấp độ thấp hơn, và vị thế đạo đức của các thực thể này thường bị chối bỏ. Chúng ta không thấy tác hại nào trong việc xóa bỏ một ứng dụng ra khỏi điện thoại thông minh, và chúng ta không nghĩ rằng một nhà phẫu thuật thần kinh không làm hại ai khi tách một module lỗi ra khỏi một bộ não của người động kinh. Còn đối với những tổ hợp được tổ chức đặc biệt ở cấp độ tương đương não người, có lẽ hầu hết chúng ta sẽ coi chúng có giá trị đạo đức quan trọng nếu chúng ta cho rằng chúng có năng lực hay tiềm năng trải nghiệm nhận thức.²⁶

Vì vậy chúng ta có thể hình dung, như trong một trường hợp cực đoan, một xã hội có trình độ phát triển công nghệ cao, gồm nhiều cấu trúc phức tạp, một vài trong số chúng thông minh và phức tạp hơn nhiều so với bất cứ thứ gì hiện đang tồn tại trên hành tinh của chúng ta nhưng lại thiếu vắng bất cứ loại hình sống nào có nhận thức hay

phúc lợi với một tầm quan trọng nhất định. Theo một nghĩa nào đó, đây có thể được coi là một xã hội không có người. Đó có thể là một xã hội của những phép màu kinh tế và những sự phát triển công nghệ đáng kinh ngạc nhưng không có ai để hưởng lợi. Một công viên Disney vắng bóng trẻ em.

Tiến hóa không hẳn là đi lên

Từ “tiến hóa” đôi khi được dùng như từ đồng nghĩa với “tiến bộ”, có lẽ để phản ánh một hình ảnh phổ biến ít bị phê phán của tiến hóa với tư cách là một “lực” tích cực. Một niềm tin sai chỗ đặt vào lợi ích nội tại của quá trình tiến hóa có thể cản trở sự đánh giá công bằng với tính hấp dẫn của một kết quả đa cực, trong đó tương lai của sự sống có trí tuệ được xác định bởi các động lực cạnh tranh. Bất cứ đánh giá nào như vậy có thể dựa trên một số quan điểm (ít nhất là quan điểm ngầm) về phân bố xác suất của những loại kiểu hình khác nhau có thể thích ứng được trong nôi xúp sự sống số hậu chuyển đổi. Trong những điều kiện tốt nhất cũng rất khó mà có được đáp án đúng và rõ ràng từ đám bày nhầy không thể tránh khỏi của sự bất định lan tỏa lên những vấn đề này: càng khó hơn nếu chúng ta nhìn sự việc một cách quá lạc quan.

Một cội nguồn khả dĩ cho niềm tin vào tiến hóa tự do là định hướng đi lên khá rõ ràng được tiến bộ tiến hóa trong quá khứ thể hiện. Khởi đầu từ những sinh vật sao chép sơ đẳng, tiến hóa đã tạo ra những cơ thể ngày càng “tiên tiến”, bao gồm những tạo vật có tâm trí, ý thức, ngôn ngữ và lập luận. Gần đây những tiến bộ văn hóa và công nghệ có chút tương đồng nhất định với tiến hóa sinh học đã cho phép loài người phát triển với tốc độ nhanh hơn. Tính theo trục thời gian địa chất cũng như lịch sử, bức tranh lớn có vẻ như cho thấy một

xu hướng thiên về các mức độ ngày càng cao của sự phức tạp, kiến thức, ý thức và tổ chức được điều phối hướng tới đạt mục tiêu: nếu không quá khắt khe, ta có thể gọi xu hướng này là “tiến bộ”.²⁷

Hình ảnh của tiến hóa như một sự tiến bộ liên tục tạo ra những hệ quả tốt lành khó mà tương ứng được với sự khổ đau lớn lao mà chúng ta nhìn thấy trong cả thế giới con người cũng như thế giới tự nhiên. Những người yêu quý các thành tựu của tiến hóa có thể làm như vậy từ phương diện mỹ học thay vì đạo đức. Mặc dù vậy, câu hỏi thích hợp không phải là tương lai nào là tương lai thú vị được viết trong tiểu thuyết khoa học viễn tưởng hay chiếu trong phim tài liệu khoa học về thiên nhiên, mà là tương lai nào là tương lai ta có thể sống tốt trong đó: hai vấn đề rất khác nhau.

Hơn nữa, chúng ta không có lý do gì để nghĩ rằng bất cứ tiến bộ nào từng diễn ra là chuyện đương nhiên xảy ra. Rất nhiều thứ có thể là may mắn. Sự chối bỏ này được hỗ trợ bởi việc một hiệu ứng lựa chọn quan sát sẽ lọc lại những bằng chứng chúng ta có thể có về thành tựu của những phát triển tiến hóa của riêng chúng ta.²⁸ Giả sử trên 99,9999% các hành tinh có sự sống xuất hiện nhưng lại bị tuyệt chủng trước khi phát triển tới điểm mà những người quan sát có trí tuệ có thể bắt đầu suy tư về nguồn gốc của mình. Chúng ta có thể kỳ vọng quan sát điều gì nếu tình huống xảy ra như vậy? Có thể chúng ta sẽ quan sát thấy gì đó như thứ chúng ta thực sự nhìn thấy. Giả thuyết cho rằng xác suất của một cuộc sống có trí tuệ tiến hóa trên một hành tinh nhất định nào đó thấp không tiên đoán rằng chúng ta sẽ thấy được chính mình trên một hành tinh nơi sự sống tuyệt chủng ngay từ giai đoạn ban đầu; nó chỉ có thể tiên đoán rằng chúng ta sẽ thấy mình trên một hành tinh có sự sống có trí tuệ đã tiến hóa, thậm

chỉ nếu những hành tinh này chỉ là một phần rất nhỏ trong tất cả các hành tinh nơi sự sống nguyên thủy tiến hóa. Lịch sử lâu dài của sự sống trên Trái đất do đó có thể hỗ trợ được nhiều cho tuyên bố rằng có xác suất cao – chưa nói đến chuyện không thể tránh khỏi – của việc những sinh vật bậc cao trên hành tinh này trỗi dậy.²⁹

Thứ ba, thậm chí nếu những điều kiện hiện tại là hoàn hảo, và kể cả nếu có thể chứng minh được rằng chúng đương nhiên phát sinh từ trạng thái sơ khai chung nào đó, thì vẫn không có gì đảm bảo rằng xu hướng mang tính cải thiện này sẽ tiếp tục kéo dài vô tận trong tương lai. Điều này đúng cả khi chúng ta bỏ qua xác suất của một sự kiện tuyệt chủng bởi thảm họa và ngay cả nếu chúng ta có giả định rằng những phát triển tiến hóa sẽ tiếp tục để tạo ra các hệ thống ngày càng phức tạp.

Chúng tôi đã gợi ý ở trên rằng những công nhân trí tuệ máy được lựa chọn bởi năng suất lao động tối đa có thể làm việc vô cùng chăm chỉ và vẫn chưa rõ những công nhân này sẽ hạnh phúc đến mức nào. Chúng tôi cũng đặt ra khả năng rằng những dạng thức thích hợp nhất trong một nôi xúp sự sống số cạnh tranh trong tương lai có thể không có ý thức. Gần như thiếu vắng hoàn toàn niềm vui hay ý thức, có thể dẫn đến sự lãng phí những phẩm chất khác có thể được coi như không thể thiếu cho một đời sống tốt. Con người quý trọng âm nhạc, hài hước, tình cảm lãng mạn, nghệ thuật, chơi đùa, nhảy múa, chuyện trò, triết học, văn học, phiêu lưu, khám phá, thức ăn và đồ uống, tình bạn, làm cha mẹ, thể thao, thiên nhiên, truyền thống và tâm linh, và nhiều thứ khác nữa. Không có gì đảm bảo rằng có bất kỳ điều gì trong số đó sẽ còn thích ứng được. Có thể thứ sẽ làm tăng cực đại sự phù hợp sẽ không phải là gì khác ngoài lao động cực

nhọc, công việc buồn tẻ và nhàm chán không ngừng nghỉ với cường độ cao, những niềm vui tẻ nhạt ít ỏi, chỉ nhằm mục tiêu cải thiện một con số nào đó đứng thứ tám sau dấu phẩy của một chỉ tiêu kinh tế đầu ra nào đó. Những kiểu hình được lựa chọn sau đó có thể có cuộc sống không mang những phẩm chất nói trên, và tùy theo thuyết giá trị của từng người mà kết quả người đó có thể cảm nhận sẽ là đáng khinh, vô giá trị hoặc chỉ đơn thuần là nghèo khó, nhưng dấu sao thì nó cũng cách rất xa một thế giới không tưởng mà người đó có thể thấy đáng để khen ngợi.

Ta có thể băn khoăn làm sao mà một bức tranh ảm đạm như vậy lại có thể nhất quán được với việc chúng ta lại đang thích thú hưởng thụ âm nhạc, hài hước, tình cảm lãng mạn,... Nếu những hành vi này thực sự là “lãng phí” thì tại sao chúng lại được chấp nhận và thực sự được khuyến khích bởi các quá trình tiến hóa hình thành nên giống loài của chúng ta? Việc con người hiện đại đang nằm trong trạng thái mất cân bằng tiến hóa không tính đến điều này; bởi vì kể cả các vị tổ tiên thời tiền sử của chúng ta cũng lại tham gia vào hầu hết những hoạt động uổng phí này. Nhiều hành vi đang được chú ý này thậm chí không phải là độc nhất ở *Homo sapiens*. Hiện tượng khoa trương (*flamboyant display*) còn xuất hiện trong những bối cảnh rộng lớn hơn, từ việc lựa chọn bạn tình trong thế giới động vật cho tới những cuộc cạnh tranh uy tín giữa các quốc gia.³⁰

Mặc dù một cách giải thích đầy đủ mang tính tiến hóa cho từng hành vi trong số đó là nằm ngoài phạm vi của nghiên cứu này, nhưng chúng ta có thể ghi nhận rằng một vài hành vi trong số đó có những chức năng có thể không phù hợp trong bối cảnh trí tuệ máy. Chơi đùa chẳng hạn, chỉ có ở một số loài và phần lớn trong các con non,

chủ yếu là một cách để con non học những kỹ năng sau này sẽ cần đến trong cuộc đời. Trong các bản giả lập có thể được tạo ra ngay như người lớn đã sẵn có những bộ kỹ năng trưởng thành hay khi kiến thức và các kỹ thuật do một AI đạt được có thể được trực tiếp chuyển sang cho AI khác, thì nhu cầu đối với hành vi chơi đùa có thể sẽ kém phổ biến hơn.

Nhiều ví dụ khác về hành vi của con người có thể đã được tiến hóa như những tín hiệu khó có thể giả mạo về những phẩm chất khó quan sát được trực tiếp, ví dụ như sự dẻo dai của cơ thể và tinh thần, địa vị xã hội, chất lượng của những người đồng minh, khả năng và mong muốn chiến thắng trong các cuộc chiến đấu hoặc sự sở hữu nguồn lực. Cái đuôi của con công đực là một ví dụ cổ điển: chỉ những con công đực thích hợp mới biết xòe ra một cái đuôi thật rực rỡ, còn những con công cái đã được tiến hóa để thấy điều đó là hấp dẫn. Không kém hơn những tổ chất hình thể, những tổ chất hành vi cũng báo hiệu sự phù hợp di truyền hay các thuộc tính xã hội liên quan khác.³¹

Cho rằng hiện tượng khoa trương là khá phổ biến ở con người và những loài khác, ta có thể xem xét liệu đó cũng có thể trở thành một phần trong dự trữ các hình thái sự sống tiên tiến hơn về công nghệ hay không. Thậm chí nếu không có ứng dụng hạn hẹp mang tính tiện ích nào cho việc chơi đùa hay cảm thụ âm nhạc, thậm chí cho cả nhận thức trong hệ sinh thái tương lai của xử lý thông tin trí tuệ, thì những tổ chất này có đem lại ưu thế tiến hóa nào đó cho chủ nhân của nó như những tín hiệu đáng tin cậy của những phẩm chất thích ứng khác hay không?

Trong khi khả năng của một sự hài hòa được hình thành từ trước giữa thứ có giá trị với chúng ta và thứ có thể thích ứng được trong một hệ sinh thái số tương lai là khó có thể loại bỏ, vẫn có lý do để hoài nghi. Trước hết hãy xem xét rằng nhiều hình thức thể hiện đất đỏ mà chúng ta thấy trong tự nhiên thường gắn với việc lựa chọn bạn tình.³² Sự sinh sản trong các hình thái sự sống trưởng thành về mặt công nghệ, ngược lại, có thể phần lớn hoặc hoàn toàn phi tính dục.

Thứ hai, các đại diện tiên tiến về công nghệ có thể có những phương pháp mới sẵn có để truyền đạt thông tin một cách đáng tin cậy với nhau, những công cụ không cần trông cậy vào phô trương đất đỏ. Thậm chí ngày nay, khi những người cho vay chuyên nghiệp đánh giá hạng tín dụng thì họ thường trông cậy nhiều hơn vào chứng từ, ví dụ như chứng nhận sở hữu hoặc sao kê ngân hàng hơn là vào những phô trương đất đỏ chẳng hạn như quần áo hàng hiệu hay đồng hồ Rolex. Trong tương lai, có thể thuê các công ty kiểm toán để xác minh thông qua việc kiểm tra kỹ lưỡng lịch sử hành vi, kiểm định trong những môi trường mô phỏng hoặc thanh tra trực tiếp mã nguồn, rằng một đại diện khách sở hữu một thuộc tính như đã tuyên bố. Báo hiệu về những phẩm chất của ai đó bằng cách nhất trí với sự kiểm toán này có thể hiệu quả hơn là báo hiệu thông qua những biểu hiện khoa trương. Một tín hiệu được tạo ra một cách chuyên nghiệp như vậy có thể khá đắt để làm giả – đó là một tính năng quan trọng giúp tín hiệu trở nên đáng tin cậy – nhưng có thể sẽ rẻ hơn nhiều để truyền đi khi nó đúng sự thật so với việc truyền đạt tín hiệu tương đương một cách khoa trương.

Thứ ba, không phải tất cả các kiểu phát tín hiệu đất đỏ đều có giá trị nội tại hay được xã hội mong muốn. Nhiều thứ chỉ đơn giản là phí phạm. Những nghi lễ tặng quà của người Kwakiuti, một dạng thi thố về địa vị giữa các tù trưởng thù địch, bao gồm việc phá hoại một lượng lớn tài sản đã được tích lũy.³³ Những tòa nhà chọc trời kỳ lục, các siêu du thuyền và tên lửa du hành cung trăng có thể được xem như những ví dụ hiện đại. Trong khi những hoạt động như âm nhạc hoặc hài hước có thể được coi như làm tăng thêm giá trị nội tại của cuộc sống con người, thì chưa chắc một tuyên bố tương tự có thể đúng với việc theo đuổi những phụ kiện trang phục đất đỏ hay những biểu tượng địa vị khác của người tiêu dùng. Tệ hại hơn nữa, kiểu phát tín hiệu đất đỏ có thể còn gây hại, như việc thể hiện nam tính dẫn đến bạo lực băng đảng hoặc dũng cảm trong quân ngũ. Thậm chí nếu các dạng thức sự sống tương lai có thể sử dụng kiểu phát tín hiệu đất đỏ, thì liệu tín hiệu đó có giá trị gì hay không có thể vẫn còn là một câu hỏi mở – liệu nó có giống như nhịp điệu tươi vui của một con họa mi hay thay vào đó lại là tiếng nghiêng răng đơn điệu của cóc (hoặc tiếng sữa dai dẳng của một con chó mắc bệnh dại)?

Sự hình thành hậu chuyển đổi của một thể đơn nhất

Thậm chí nếu kết quả trung gian của việc chuyển đổi sang trí tuệ máy là đa cực, vẫn có khả năng một thể đơn nhất sẽ tiếp tục phát triển sau này. Sự phát triển này có thể tuân theo một xu hướng dài hạn rõ ràng hướng tới những quy mô lớn hơn của hội nhập chính trị, đưa nó vào kết luận tự nhiên cuối cùng của mình.³⁴ Việc đó có thể diễn ra như thế nào?

Lần chuyển đổi thứ hai

Có một cách để một kết quả đa cực ban đầu hội tụ thành một thể đơn nhất hậu chuyển đổi: nếu như sau chuyển đổi ban đầu có một chuyển đổi công nghệ thứ hai đủ lớn và đủ nhanh để đem lại ưu thế chiến lược quyết định cho một trong những “thế lực” còn lại, mà sau đó có thể nắm bắt cơ hội để thiết lập một thể đơn nhất. Sự chuyển đổi giả thuyết thứ hai này có thể xảy ra nhờ một đột phá lên cấp độ cao hơn của siêu trí tuệ. Ví dụ, nếu làn sóng đầu tiên của siêu trí tuệ máy dựa trên giả lập thì sự trỗi dậy thứ hai có thể xảy ra khi các bản giả lập đang tiến hành nghiên cứu phát triển thành công trí tuệ nhân tạo có khả năng tự cải thiện hiệu quả.³⁵ (Ngoài ra, một chuyển đổi thứ hai có thể xảy ra do một đột phá trong công nghệ nano hay một vài công nghệ quân sự hay tổng quát khác hiện vẫn chưa được hình dung ra.)

Tốc độ phát triển sau lần chuyển đổi ban đầu có thể là cực kỳ nhanh. Thậm chí một khoảng cách ngắn giữa thế lực đi đầu và đối thủ cạnh tranh vì thế có thể sẽ tạo ra một ưu thế chiến lược quyết định cho thế lực dẫn đầu trong lần chuyển đổi thứ hai. Giả sử hai dự án bước vào chuyển đổi lần thứ nhất chỉ cách nhau vài ngày, và việc cất cánh đủ chậm sao cho khoảng cách không đem lại cho dự án dẫn đầu một ưu thế chiến lược quyết định vào bất cứ thời điểm nào trong khi cất cánh. Cả hai dự án đều nổi lên như những thế lực siêu trí tuệ, mặc dù một trong hai vẫn đi trước thế lực kia vài ngày. Nhưng những diễn biến đang xảy ra trong trục thời gian nghiên cứu đặc trưng cho siêu trí tuệ máy – có thể hàng ngàn hay hàng triệu lần nhanh hơn nghiên cứu được thực hiện trên trục thời gian của con người sinh học. Sự phát triển của công nghệ trong lần chuyển đổi thứ hai vì thế có thể được hoàn thành chỉ trong vài ngày, vài giờ hay vài phút.

Thậm chí ngay cả khi khoảng cách của dự án dẫn đầu chỉ là vài ngày, một đột phá có thể phóng nó tới một ưu thế chiến lược quyết định. Tuy nhiên hãy lưu ý rằng nếu việc khuếch tán công nghệ (thông qua gián điệp hoặc các kênh khác) tăng tốc nhanh như phát triển công nghệ thì hiệu ứng này có thể mất tác dụng. Vậy điều còn liên quan tiếp theo là độ dốc của chuyển đổi lần hai, là tốc độ diễn ra của nó so với tốc độ chung của các sự kiện trong giai đoạn sau chuyển đổi lần đầu. (Theo nghĩa này, sự việc diễn ra càng nhanh sau chuyển đổi lần đầu thì độ dốc của chuyển đổi lần hai càng nhỏ.)

Ta có thể còn giả sử rằng một lợi thế cạnh tranh quyết định có thể dễ được thực sự sử dụng để hình thành một thể đơn nhất hơn nếu nó xuất hiện trong chuyển đổi thứ hai (và tiếp theo). Sau lần chuyển đổi thứ nhất, những người ra quyết định sẽ có thể là siêu trí tuệ hoặc tiếp cận được với tư vấn từ một siêu trí tuệ, tư vấn này có thể làm rõ ảnh hưởng của các lựa chọn chiến lược sẵn có. Hơn nữa tình hình sau chuyển đổi lần một có thể là một động thái phòng ngừa chống lại những đối thủ cạnh tranh tiềm tàng có thể ít nguy hiểm hơn đối với kẻ ra tay. Nếu các trí óc ra quyết định sau chuyển đổi lần một là trí óc số, chúng có thể được sao chép và vì vậy trở nên ít bị tổn thương hơn bởi một cuộc phản công. Thậm chí nếu một người bảo vệ có khả năng giết chết 90% dân số của bên ra tay trong một cuộc trả đũa thì điều đó cũng gần như không hữu dụng khi những kẻ bị tiêu diệt có thể được phục sinh ngay tức khắc từ các bản sao dự phòng. Phá hủy cơ sở hạ tầng (có thể xây dựng lại) có thể coi là chấp nhận được đối với các trí óc số có tuổi thọ vô hạn, những người có thể lên kế hoạch tăng cực đại nguồn lực và ảnh hưởng lên thang thời gian vũ trụ.

Các siêu sinh vật và các nền kinh tế khổng lồ

Kích thước của những tổ chức do con người điều phối, ví dụ như các công ty hay quốc gia, bị ảnh hưởng bởi những tham số khác nhau – công nghệ, quân sự, tài chính và văn hóa – những thứ có thể khác nhau tùy theo từng giai đoạn lịch sử. Một cuộc cách mạng trí tuệ máy có thể đòi hỏi những thay đổi cơ bản ở nhiều trong số các tham số này. Có thể những thay đổi đó có thể thúc đẩy sự trỗi dậy của một thể đơn nhất. Mặc dù nếu không đi vào chi tiết những thay đổi tiềm tàng này, chúng ta không thể loại trừ xác suất ngược lại (rằng những thay đổi này sẽ khuyến khích sự phân mảnh thay vì hợp nhất), nhưng vẫn có thể ghi nhận rằng chính sự khác biệt và bất định gia tăng mà chúng ta đang phải đối mặt ở đây có thể trở thành cơ sở để tin tưởng hơn vào sự xuất hiện khả thể của một thể đơn nhất so với bình thường. Một cuộc cách mạng trí tuệ máy có thể xáo trộn mọi thứ – có thể xào lại cỗ bài để đem lại những điều chỉnh có thể về địa chính trị mà có lẽ chưa từng có trong các quân bài.

Một phân tích toàn diện về tất cả những yếu tố có thể ảnh hưởng đến hội nhập chính trị có thể đưa chúng ta rời xa phạm vi của cuốn sách này: chúng ta có thể viết cả một bộ sách chỉ bằng cách rà soát các tài liệu kinh tế và chính trị học. Chúng tôi phải tự giới hạn mình vào việc đưa ra những ám chỉ ngắn gọn tới một vài yếu tố, khía cạnh của việc tối ưu hóa các tác tử có thể giúp cho việc tập trung kiểm soát được dễ dàng hơn.

Carl Shulman lập luận rằng trong một số đông các bản giả lập, những áp lực chọn lọc có thể thiên vị sự nổi lên của “các siêu sinh vật”, các nhóm giả lập sẵn sàng tự hy sinh vì lợi ích của cộng đồng mình.³⁶

Các siêu sinh vật có thể không gặp phải các vấn đề đại diện đang bủa vây các tổ chức có thành viên đang theo đuổi lợi ích cá nhân của riêng mình. Như các tế bào trong cơ thể chúng ta, hoặc những con vật riêng lẻ trong đàn kiến có tập tính xã hội cao, các bản giả lập hoàn toàn vị tha trong quan hệ với những người anh em sao chép của mình có thể cộng tác với nhau ngay cả khi không có những hình thức tưởng thưởng khéo léo.

Các siêu sinh vật có một ưu thế đặc biệt mạnh mẽ nếu việc xóa bỏ phi đồng thuận (hoặc đình chỉ vô thời hạn) các bản giả lập riêng lẻ là không được phép. Các công ty hay các quốc gia sử dụng các bản giả lập đòi hỏi quyền tự bảo tồn có thể phải mang cái ách cam kết vĩnh viễn trả tiền bảo trì cho những công nhân hết thời hoặc dư thừa. Ngược lại, các tổ chức sở hữu các bản giả lập sẵn sàng tự xóa khi không còn được cần đến có thể dễ dàng thích ứng với sự biến đổi của nhu cầu; và họ có thể tự do thử nghiệm, sản sinh ra nhiều biến thể công nhân khác nhau và chỉ giữ lại những biến thể có năng suất cao nhất.

Nếu việc xóa bỏ không tự nguyện không phải không được phép thì ưu thế so sánh của các bản giả lập có tổ chức cao giảm xuống, mặc dù không phải là bị loại bỏ hoàn toàn. Những người thuê mướn các cá thể tự hy sinh, hoạt động cộng tác có thể tăng được hiệu năng từ các vấn đề đại diện nhỏ hơn trong toàn tổ chức, bao gồm cả việc tránh được rắc rối xuất phát từ hành động đàn áp bất cứ sự chống đối nào của các bản giả lập không muốn bị xóa bỏ. Nhìn chung, tăng trưởng năng suất từ việc công nhân sẵn sàng hy sinh cuộc sống riêng của mình cho tài sản chung là một trường hợp đặc biệt của những lợi ích mà một tổ chức có thể nhận được từ việc có những

thành viên tận tụy đến mức mê muội với nó. Những thành viên này có thể không chỉ lao xuống mồ vì tổ chức và làm việc cần mẫn không cần tiền lương: họ còn tránh xa chính trị công sở và cố gắng hành động một cách nhất quán vì điều mà họ tin là lợi ích tốt nhất của tổ chức, giảm thiểu nhu cầu giám sát và những rào cản quan liêu.

Nếu như cách duy nhất để có được sự tận tụy này là hạn chế tư cách thành viên đối với các bản sao anh em (sao cho tất cả các bản giả lập trong một siêu cơ thể cụ thể được in ra từ cùng một bản mẫu) thì các siêu cá thể chịu một số bất lợi khi chỉ có thể khai thác được một số lượng kỹ năng ít hơn so với các tổ chức đối nghịch, một bất lợi có thể hoặc không đủ lớn để vượt qua được những ưu điểm của việc tránh được các vấn đề nội bộ của đại diện.³⁷ Nhược điểm này có thể được giảm thiểu đáng kể nếu một siêu cơ thể có thể ít nhất bao gồm những thành phần được huấn luyện khác biệt. Thậm chí nếu tất cả thành phần của nó được tạo ra từ một khuôn mẫu duy nhất thì lực lượng lao động của nó vẫn có thể đóng góp những kỹ năng đa dạng. Bắt đầu với một mẫu giả lập có tài năng tổng hợp, các hậu duệ có thể tỏa nhánh sang những chương trình huấn luyện khác nhau, một bản sao học kế toán, một bản khác học kỹ thuật điện... Việc này sẽ tạo ra một hội với những kỹ năng đa dạng mặc dù không phải những tài năng đa dạng. (Sự đa dạng tối đa có thể đòi hỏi sử dụng nhiều khuôn mẫu.)

Thuộc tính cốt yếu của một siêu cơ thể không phải là nó được cấu thành bởi các bản sao của một tổ tiên duy nhất mà là tất cả các tác tử riêng biệt bên trong nó cam kết hoàn toàn với một mục tiêu chung. Khả năng tạo ra một siêu cơ thể vì thế có thể được xem như đòi hỏi một giải pháp từng phần cho vấn đề kiểm soát. Vì một giải

pháp chung đầy đủ cho vấn đề kiểm soát có thể sẽ cho phép ai đó tạo ra một tác tử với mục tiêu tối thượng hoàn toàn ngẫu nhiên, giải pháp một phần cần để tạo ra một siêu cơ thể chỉ đòi hỏi năng lực trao cho nhiều tác tử mục tiêu tối thượng như nhau (cho một số mục tiêu tối thượng không tầm thường nhưng không nhất thiết ngẫu nhiên).³⁸

Quan điểm chính được đưa ra trong phần này vì thế không thực sự bị hạn chế trong các nhóm giả lập nhân bản đơn tính, mà còn có thể được phát biểu khái quát hơn để làm rõ rằng nó áp dụng cho nhiều kịch bản siêu trí tuệ máy đa cực. Rằng những tiến bộ nhất định của các kỹ thuật lựa chọn khuyến khích (những tiến bộ có thể trở nên khả thi khi những đối tượng hành động là đối tượng số) có thể giúp vượt qua được một số hoạt động kém hiệu quả hiện đang cản trở các tổ chức lớn của con người và đang làm đối trọng của kinh tế quy mô. Khi những giới hạn này được dỡ bỏ, các tổ chức – bất kể là các công ty, các quốc gia hay các thực thể kinh tế hoặc chính trị khác – có thể tăng kích cỡ. Đó là một yếu tố có thể thúc đẩy sự nổi lên của một thể đơn nhất hậu chuyển đổi.

Một phạm trù mà trong đó, các siêu cơ thể (hoặc các tác tử số khác với những động lực được chọn lọc một phần) có thể vượt trội là chế độ áp bức. Một quốc gia có thể sử dụng phương pháp chọn lọc tạo động lực để bảo đảm rằng các lực lượng cảnh sát, quân đội, mật vụ và hành chính dân sự nhất loạt trung thành. Như Shulman ghi nhận:

Các trạng thái được lưu lại [của bản giả lập trung thành nào đó mà đã được chuẩn bị và xác minh cẩn thận] có thể được sao chép hàng tỷ lần để biên chế cho các lực lượng quân đội, giới quan liêu

hay cảnh sát đồng nhất về ý thức hệ. Sau một thời gian làm việc ngắn, mỗi bản sao sẽ được thay thế bằng một bản sao hoàn toàn mới của cùng một trạng thái đã được lưu để ngăn chặn việc lệch lạc ý thức hệ. Trong một vùng quyền hạn đã cho, năng lực này có thể cho phép giám sát và quản lý vô cùng chi tiết: chỉ có một bản sao như vậy cho mỗi người dân. Nó có thể được sử dụng để cấm phát triển vũ khí hủy diệt hàng loạt, thực thi các quy định về thí nghiệm giả lập não bộ hay sinh sản, thực thi một hiến pháp tự do dân chủ hoặc tạo ra một chủ nghĩa toàn trị kinh hoàng và vĩnh viễn.³⁹

Hiệu ứng cấp một của năng lực như vậy xem ra là củng cố và có thể là tập trung quyền lực vào một nhóm nhỏ hơn.

Thống nhất bằng hiệp ước

Có thể nhận được những lợi ích tiềm năng to lớn từ cộng tác quốc tế trong một thế giới đa cực hậu chuyển đổi. Có thể tránh được các cuộc chiến tranh và chạy đua vũ trang. Những nguồn lực từ thiên thể có thể được chiếm lĩnh và thu hoạch với tốc độ tối ưu hóa toàn cầu. Sự phát triển những dạng thức trí tuệ máy tiên tiến hơn có thể được điều phối để tránh vội vàng và để cho phép những thiết kế mới được kiểm định kỹ càng. Những phát triển khác có khả năng gây ra những rủi ro diệt vong có thể được tạm dừng. Và các quy định thống nhất có thể được thực thi trên toàn cầu, bao gồm cung cấp một tiêu chuẩn cuộc sống được bảo đảm (mà có thể đòi hỏi một hình thức kiểm soát dân số nào đó) và để ngăn chặn việc bóc lột, lạm dụng các bản giả lập cùng các loại trí óc kỹ thuật số và sinh học khác. Hơn nữa, các tác tử với các ưu tiên đáp ứng được về mặt tài nguyên (vấn đề này sẽ còn được đề cập trong chương 13) có thể đưa ra một thỏa thuận chia sẻ đảm bảo rằng chúng có được một chỗ đứng nào đó

trong tương lai, hơn là một cuộc đấu tranh được ăn cả ngã về không mà chúng có thể không nhận được gì.

Tuy nhiên, sự hiện diện của những lợi ích tiềm năng to lớn từ cộng tác không hàm ý rằng cộng tác có thể thực sự đạt được. Trong thế giới ngày nay, có thể có được nhiều món lợi từ việc điều phối toàn cầu tốt hơn, chẳng hạn như việc giảm thiểu chi phí quân sự, chiến tranh, đánh bắt hải sản quá mức, các rào cản thương mại hay ô nhiễm khí quyển. Mặc dù vậy, những quả ngọt này lại bị để hỏng trên cành. Sao lại như vậy? Điều gì ngăn chặn một kết quả hợp tác đầy đủ có thể tăng cực đại lợi ích chung?

Khó khăn trong việc đảm bảo tuân thủ với bất cứ hiệp ước nào đã được thỏa thuận, bao gồm cả các chi phí giám sát và thực thi là một cản trở. Hai đối thủ hạt nhân có thể sẽ đều có lợi nếu cả hai cùng tiêu hủy bom nguyên tử của mình; nhưng cho dù họ có đạt được thỏa thuận trên nguyên tắc để làm việc này thì việc giải trừ quân bị mặc dù vậy vẫn rất khó khăn vì nỗi sợ chung rằng bên kia có thể lừa dối. Có thể có các thanh tra để giám sát việc tiêu hủy kho vũ khí hiện có, rồi giám sát các lò phản ứng nguyên tử và các cơ sở khác, và thu thập thông tin tình báo về kỹ thuật và con người nhằm mục tiêu bảo đảm rằng chương trình vũ khí không được tái lập. Một mặt, cần chi phí để thanh toán cho những thanh tra này. Mặt khác, cũng có rủi ro rằng các thanh tra có thể làm gián điệp và bỏ trốn với các bí mật thương mại hoặc quân sự. Có lẽ đáng kể nhất là mỗi bên có thể lo sợ rằng bên kia sẽ vẫn duy trì năng lực hạt nhân bí mật. Nhiều hợp đồng có lợi ích tiềm tàng đáng kể có thể không bao giờ được thực hiện vì việc kiểm định sự tuân thủ là quá khó khăn.

Nếu có được những công nghệ thanh tra mới làm giảm chi phí giám sát, ta có thể kỳ vọng rằng điều đó sẽ giúp gia tăng mức độ hợp tác. Vẫn chưa hoàn toàn rõ liệu chi phí giám sát có thể giảm xuống hay không trong kỷ nguyên hậu chuyển đổi. Trong khi chắc chắn rằng sẽ có nhiều kỹ thuật thanh tra mạnh mẽ mới, cũng có thể sẽ xuất hiện những biện pháp giấu diếm mới. Cụ thể, một phần ngày càng lớn của các hoạt động ta muốn quản lý có thể sẽ được thực hiện trên mạng, ngoài tầm với của hoạt động giám sát vật lý. Ví dụ, các trí óc kỹ thuật số nghiên cứu thiết kế một hệ thống vũ khí công nghệ nano mới hoặc thể hệ trí tuệ nhân tạo mới có thể làm điều này mà không để lại nhiều dấu vết vật lý. Điều tra số có thể không thâm nhập được qua tất cả các lớp che giấu và mã hóa trong đó một kẻ vi phạm hiệp ước có thể che phủ các hoạt động phi pháp của mình.

Công nghệ phát hiện nói dối đáng tin cậy, nếu như có thể được phát triển, sẽ là một công cụ cực kỳ hữu ích để giám sát quá trình tuân thủ.⁴⁰ Một nghị định thư về thanh tra có thể bao gồm những điều khoản về việc phỏng vấn các quan chức chủ chốt để xác nhận rằng họ có dự kiến triển khai tất cả các điều khoản của hiệp ước và rằng họ không biết bất cứ vi phạm nào cho dù đã rất nỗ lực tìm kiếm.

Một người ra quyết định đang lên kế hoạch lừa đảo có thể đánh bại được hình thức xác minh thông qua công nghệ phát hiện nói dối này bằng cách: trước hết ra lệnh cho cấp dưới thực hiện hoạt động phi pháp và giấu hoạt động này đối với chính người ra quyết định, sau đó thực hiện một thủ tục nào đó để xóa bỏ ký ức của chính mình về việc có liên quan đến những âm mưu này. Các hoạt động xóa bỏ ký ức được định hướng thích hợp là hoàn toàn khả thi trong những bộ não sinh học với công nghệ thần kinh tiên tiến hơn. Điều này có

thể còn dễ dàng hơn trong các trí tuệ máy (tùy thuộc vào kiến trúc của chúng.)

Các nhà nước có thể tìm cách giải quyết vấn đề này bằng cách tự cam kết với một hình thức giám sát liên tục theo đó thường xuyên kiểm tra các quan chức chủ chốt bằng máy phát hiện nói dối để xem họ có bất cứ dự định phá hoại hay đánh lừa bất cứ hiệp ước nào đã được nhà nước tham gia hay có thể tham gia trong tương lai hay không. Một cam kết như vậy có thể được xem như một loại hình siêu hiệp ước, có thể thúc đẩy việc xác minh những hiệp ước khác; nhưng các nhà nước có thể đơn phương tự cam kết với nó để có được lợi ích của việc được coi là một đối tác đàm phán đáng tin. Tuy nhiên, cam kết hay siêu hiệp ước này có thể đối diện với cùng vấn đề phá hoại bằng kế giao phó- và-quên này. Một cách lý tưởng, một siêu hiệp ước có thể có hiệu lực trước khi bất cứ bên nào có cơ hội thực hiện bố trí nội bộ cần thiết để phá hoại việc triển khai thực hiện nó. Một khi những hành vi xấu xa này nắm được một thời điểm mất cảnh giác để chôn xuống những quả mìn lừa dối, lòng tin sẽ không thể tiếp tục nảy mầm.

Trong một số trường hợp, chỉ cần năng lực phát hiện là đã đủ để thiết lập sự tự tin cần thiết cho hợp đồng. Tuy nhiên, trong những trường hợp khác, lại cần có một cơ chế nào đó để cưỡng chế tuân thủ hoặc đưa ra hình phạt với các hành vi vi phạm. Nhu cầu đối với một cơ chế cưỡng chế như vậy có thể nảy sinh nếu lời đe dọa rút bên sai phạm ra khỏi hiệp ước là không đủ để ngăn chặn vi phạm, chẳng hạn như khi bên vi phạm hiệp ước có thể có được một ưu thế sao cho anh ta sau đó không cần quan tâm đến phản ứng của bên kia.

Nếu các phương pháp lựa chọn động lực hiệu quả cao là sẵn có, vấn đề cưỡng chế này có thể được giải quyết bằng cách trao cho một đại diện độc lập một lực lượng quân đội hoặc cảnh sát đủ mạnh để cưỡng chế thi hành hiệp ước ngay cả khi có sự chống đối của một hoặc nhiều bên tham gia. Giải pháp này yêu cầu bên cưỡng chế phải là bên có thể tin cậy được. Nhưng với các kỹ thuật lựa chọn động lực đủ tốt, có thể có được sự tự tin cần thiết bằng cách để tất cả các bên tham gia hiệp ước cùng giám sát quá trình thiết kế đại diện cưỡng chế này.

Chuyển giao quyền lực cho một đại diện cưỡng chế bên ngoài làm phát sinh nhiều vấn đề tương tự như những gì chúng ta đã đối mặt trước đây trong những cuộc thảo luận của chúng ta về một kết quả đa cực (kết quả trong đó một thể đơn nhất sinh ra trước hoặc trong cuộc cách mạng trí tuệ máy ban đầu). Để có thể thực thi các hiệp ước liên quan đến những lợi ích an ninh sống còn của các quốc gia đối địch, đại diện cưỡng chế bên ngoài về cơ bản phải là một thể đơn nhất: một con thủy quái siêu trí tuệ toàn cầu. Tuy nhiên, có một sự khác biệt là hiện nay chúng ta đang nghiên cứu tình huống hậu chuyển đổi, trong đó các tác tử có thể tạo ra con thủy quái này có thể có năng lực cao hơn con người chúng ta. Bản thân những kẻ sản sinh quái vật này cũng có thể đã là siêu trí tuệ. Điều đó sẽ tăng đáng kể khả năng của việc chúng có thể giải quyết được vấn đề kiểm soát và thiết kế một đại diện cưỡng chế có thể phục vụ được lợi ích của tất cả các bên có quyền tham gia chế tạo nó.

Ngoài chi phí giám sát và cưỡng chế tuân thủ, liệu có thể còn có chướng ngại nào khác đối với điều phối toàn cầu hay không? Có thể vấn đề chính còn lại là điều mà chúng ta có thể tham chiếu đến như

chi phí thương lượng.⁴¹ Ngay cả khi có thể có một phương án tốt đem lại lợi ích cho tất cả, nó cũng có thể không bao giờ thành hiện thực vì các bên không thống nhất được cách phân chia chiến lợi phẩm. Ví dụ, nếu hai người ký một hợp đồng có thể đem lại một đô la lợi nhuận nhưng mỗi bên đều thấy rằng mình xứng đáng được nhận 60 cent và từ chối nhận ít hơn, hợp đồng sẽ không được thực hiện và lợi ích tiềm tàng sẽ bị mất đi. Nhìn chung, các cuộc thảo luận có thể khó khăn hoặc kéo dài, hoặc hoàn toàn không có kết quả vì những lựa chọn thương lượng chiến lược do một số bên đưa ra.

Trong cuộc đời thực, con người thường thành công trong việc đạt được thỏa thuận cho dù có xác suất của thương lượng chiến lược (mặc dù vẫn phải tốn khá nhiều thời gian và sự kiên nhẫn). Tuy nhiên, có thể hiểu được rằng các vấn đề thương lượng chiến lược có thể có một động lực khác trong kỷ nguyên hậu chuyển đổi. Một nhà đàm phán AI có thể bám theo một số quan điểm chính thống cụ thể nào đó về sự hợp lý một cách nhất quán hơn, có thể với những hệ quả mới hoặc bất ngờ khi đối chiếu với các nhà đàm phán AI khác. Một AI có thể có những nước đi trong ván cờ thương lượng mà con người không có hoặc rất khó thực hiện, bao gồm khả năng cam kết trước với một chính sách hay một quá trình hành động. Trong khi con người (và các thể chế do con người điều hành) không mấy khi có thể cam kết trước – mà nếu có thì cũng không hoàn toàn đáng tin cậy và cụ thể – thì một số loại trí tuệ máy lại có thể đưa ra những lời cam kết trước không thể phá vỡ và cho phép các đối tác đàm phán xác nhận rằng lời cam kết như vậy đã được đưa ra.⁴²

Tính sẵn có của các kỹ thuật cam kết trước mạnh mẽ có thể thay đổi về cơ bản chất của các cuộc đàm phán, có tiềm năng đem lại ưu

thế rất lớn cho các tác tử sở hữu lợi thế của người đi trước. Nếu sự tham gia của một tác tử cụ thể là cần thiết để hiện thực hóa lợi ích tiềm tàng nào đó có được từ hợp tác, và nếu tác tử đó có thể đi nước đầu tiên, thì nó có thể sẽ có khả năng quyết định việc phân chia chiến lợi phẩm bằng cách cam kết trước không chấp nhận bất cứ thỏa thuận nào đem lại ít hơn, ví dụ, 99% giá trị thắng dư. Những tác tử khác khi đó sẽ phải đối mặt với lựa chọn giữa việc không có được gì (khi từ chối đề nghị không công bằng) hoặc nhận 1% giá trị (khi chấp nhận). Nếu cam kết trước của tác tử đi đầu có thể xác minh công khai, các đối tác đàm phán của nó có thể chắc chắn rằng mình chỉ có hai lựa chọn này.

Để tránh bị bóc lột theo cách này, các tác tử có thể cam kết trước với việc từ chối tổng tiền hoặc tất cả các đề xuất không công bằng. Khi một cam kết trước như vậy đã được đưa ra (và được công khai hóa thành công) thì những tác tử khác có thể thấy không có ích lợi gì trong việc đưa ra các đe dọa hoặc tự cam kết trước với việc chỉ chấp nhận các thỏa thuận có lợi cho riêng mình, vì họ biết rằng các mối đe dọa sẽ thất bại và rằng những đề xuất không công bằng sẽ bị từ chối. Nhưng điều đó chỉ cho thấy một lần nữa rằng những người đi đầu sẽ có ưu thế. Tác tử đi nước đi đầu tiên có thể lựa chọn đặt cược vị thế sức mạnh của mình chỉ để ngăn chặn những tác tử khác lấy được ưu thế không công bằng hoặc chiếm được phần lớn chiến lợi phẩm tương lai.

Có lẽ tác tử có điều kiện thuận lợi nhất là tác tử khởi đầu với một tính cách hay một hệ giá trị khiến nó không bị tác động bởi sự cưỡng đoạt hoặc bất cứ thỏa thuận nào mà trong đó, sự tham gia của nó là bắt buộc nhưng nó lại không nhận được phần nhiều lợi ích. Một số

người có vẻ đã sở hữu những tố chất cá nhân tương ứng với những khía cạnh khác nhau của một tinh thần không nhân nhượng.⁴³ Tuy nhiên, một thiên hướng dễ bị kích động có thể sẽ có tác động ngược nếu bên ngoài còn có những tác tử khác cũng thấy rằng họ có quyền được nhận nhiều hơn phần công bằng của mình và cam kết không nhượng bộ. Lực không thể chặn có thể gặp vật không thể dời, khiến cho không thể đạt được thỏa thuận (hoặc tệ hại hơn: chiến tranh toàn diện). Những kẻ nhu mì yếu đuối ít nhất sẽ nhận được gì đó, mặc dù ít hơn phần chia công bằng của họ.

Chưa rõ sự cân bằng lý thuyết trong trò chơi nào có thể đạt được trong trò chơi thương lượng hậu chuyển đổi này. Các tác tử có thể lựa chọn những chiến lược phức tạp hơn nữa. Chúng ta hy vọng về việc đạt được một điểm cân bằng xoay quanh một tiêu chuẩn công tâm nào đó đóng vai trò như một điểm Schelling – một tính năng nổi bật trong một không gian kết quả lớn mà, vì những kỳ vọng chung, có thể trở thành điểm điều phối khả thể trong một cuộc chơi điều phối chưa ngã ngũ nếu không có điểm này. Một sự cân bằng như vậy có thể được bóc lột bởi một số tính cách đã tiến hóa và lập trình văn hóa nào đó của chúng ta: giả sử chúng ta thành công trong việc đưa các giá trị của mình vào kỷ nguyên hậu chuyển đổi, việc sự cân bằng trở thành mối quan tâm chung có thể làm thiên lệch các kỳ vọng và chiến lược theo những cách thức có thể dẫn đến một sự cân bằng hấp dẫn.⁴⁴

Trong mọi trường hợp, điểm tốt là với khả năng của các dạng thức cam kết trước mạnh mẽ và linh hoạt, các kết quả đàm phán có thể mang vẻ ngoài không quen thuộc. Nếu kỷ nguyên hậu chuyển đổi bắt đầu dưới dạng đa cực, một thể đơn nhất vẫn có thể xuất hiện

gần như ngay tức khắc như là hệ quả của một hiệp ước được đàm phán để giải quyết tất cả các vấn đề phối hợp quan trọng toàn cầu. Một vài chi phí giao dịch, có thể bao gồm cả chi phí giám sát và thực thi, có thể giảm mạnh cùng với những năng lực công nghệ sẵn có của các trí tuệ máy tiên tiến. Những chi phí khác liên quan đến đàm phán chiến lược, có thể vẫn duy trì ở mức cao. Nhưng bất kể đàm phán chiến lược ảnh hưởng đến bản chất của thỏa thuận đã đạt được ra sao, thì vẫn không có lý do rõ ràng cho việc nó sẽ trì hoãn quá trình tiến tới một thỏa thuận khác. Nếu không có thỏa thuận nào thành công, xung đột ở một dạng nào đó hẳn sẽ xảy ra; và có thể một bên sẽ thắng và hình thành một thể đơn nhất xung quanh liên minh chiến thắng, hoặc kết quả có thể là một cuộc xung đột không có hồi kết. Trong trường hợp thứ hai, một thể đơn nhất có thể không bao giờ được hình thành và kết quả chung có thể kém hơn rất nhiều so với những gì nó có thể sẽ và phải đạt được nếu loài người và hậu duệ của nó hành động theo phương thức có tính phối hợp hơn.

Chúng ta đã thấy rằng sự đa cực, thậm chí nếu có thể đạt được dưới dạng ổn định, có thể không bảo đảm một kết quả hấp dẫn. Vấn đề ủy thác-đại diện ban đầu vẫn chưa được giải quyết và chôn lấp nó dưới một tập những vấn đề mới liên quan đến những sai lầm điều phối chung hậu chuyển đổi chỉ có thể khiến tình hình thêm tệ hại. Do đó hãy để chúng tôi trở lại câu hỏi làm thế nào ta có thể an toàn giữ một AI siêu trí tuệ duy nhất.

Chương 12

Thụ đắc các giá trị

Ở trạng thái tối ưu, việc kiểm soát năng lực chỉ là biện pháp tạm thời và mang tính bổ trợ. Lựa chọn động lực là thứ cần được làm chủ hoàn toàn, trừ khi chúng ta có ý định giam cầm siêu trí tuệ mãi mãi. Nhưng làm sao truyền tải được một giá trị nào đó vào một tác tử nhân tạo và biến giá trị đó trở thành mục tiêu tối thượng? Tác tử này không có trí tuệ, vì thế nó có thể không hiểu, hay thậm chí không hình dung ra được bất kỳ giá trị nào có ý nghĩa với con người. Tuy nhiên, nếu trì hoãn quá trình này cho đến khi tác tử có được siêu trí tuệ, thì có thể nỗ lực của chúng ta trong việc can thiệp vào hệ thống động lực của nó sẽ bị kháng cự – và như đã bàn đến trong Chương 7, tác tử này sẽ có được những lý do hội tụ công cụ để làm như vậy. Quả thực, vấn đề truyền tải giá trị (value-loading) này là một bài toán hóc búa, nhưng phải được giải quyết.

Vấn đề truyền tải giá trị

Chúng ta không thể tính được hết tất cả các tình huống mà một siêu trí tuệ có thể rơi vào, cũng như cách nó phản ứng với từng tình huống cụ thể. Tương tự, việc tạo lập một danh sách những thế giới khả thể tồn tại và gán cho mỗi thế giới một giá trị riêng cũng là chuyện trên trời. Ở một vương quốc phức tạp hơn nhiều lần so với trò cờ ca rô, số lượng các trạng thái khả thể (và lịch sử của chúng) là

quá lớn để liệt kê. Vì vậy, một hệ thống động lực sẽ khó có thể được xác định cụ thể như một bảng dữ liệu tra cứu hỗn hợp. Thay vào đó, nó cần được biểu diễn trừu tượng hơn, như một công thức hay quy tắc cho phép tác tử quyết định điều mình sẽ làm trong bất kỳ tình huống cho trước nào.

Một quy tắc quyết định như vậy có thể được xác định thông qua một hàm thỏa dụng. Hàm thỏa dụng (xem Chương 1) sẽ gán giá trị cho mỗi kết quả có thể đạt được, hoặc bao quát hơn, vào mỗi “thế giới khả thể”, và với hàm này, ta có thể xác định một tác tử với độ thỏa dụng mong đợi ở mức tối đa. Trong mỗi tình huống, tác tử này sẽ lựa chọn một hành động có độ thỏa dụng mong đợi cao nhất. (Độ thỏa dụng mong đợi được tính toán bằng cách tính trọng số của độ thỏa dụng cho từng thế giới khả thể với xác suất chủ quan là xác suất có điều kiện của thế giới thực trên một hành động cụ thể được thực hiện.) Trong thực tế, số lượng các kết quả thu được là quá lớn để có thể tính toán độ thỏa dụng mong đợi của một hành động nào đó. Mặc dù vậy, quy tắc quyết định và hàm thỏa dụng sẽ cùng nhau xác định một điều kiện lý tưởng danh định (hay một khái niệm về tính tối ưu). Dựa vào điều kiện này, một tác tử có thể được thiết kế tiệm cận, và mức độ tiệm cận sẽ ngày một lớn khi tác tử thông minh hơn.¹ Việc tạo ra một cỗ máy tính toán được giá trị tiệm cận phù hợp của độ thỏa dụng mong đợi cho các hành động mà nó có là một vấn đề liên quan đến phạm trù AI toàn diện.²

Tuy nhiên ở chương này, chúng ta sẽ thảo luận về một vấn đề khác, vẫn tồn tại ngay cả khi đã giải được bài toán chế tạo máy móc thông minh.

Bộ khung về một tác tử tối đa hóa độ thỏa dụng có thể được vận dụng để giải quyết tình trạng lực bất tòng tâm của một lập trình viên AI hạt giống nào đó có ý định tìm ra lời giải cho bài toán kiểm soát năng lực bằng cách tạo dựng cho AI một mục tiêu tối thượng tương ứng với một ý tưởng khả thi của con người về một kết quả xứng đáng. Lập trình viên này có trong đầu một giá trị nhân văn nhất định và muốn AI thúc đẩy giá trị đó. Ta có thể lấy niềm vui làm một ví dụ cụ thể. (Những vấn đề tương tự sẽ nảy sinh khi người này quan tâm đến công lý, tự do, vinh quang, quyền con người, dân chủ, cân bằng sinh thái hay phát triển bản thân.) Nói đến bộ khung về độ thỏa dụng mong đợi, lập trình viên này sẽ tìm kiếm một hàm thỏa dụng có chức năng gán độ thỏa dụng cho từng thế giới khả thể theo tỷ lệ thuận với định lượng niềm vui mà thế giới đó chứa đựng. Nhưng anh ta sẽ thể hiện hàm thỏa dụng này bằng mã máy tính ra sao? Các loại ngôn ngữ máy tính không sở hữu dữ liệu nguyên thủy là những cụm từ như “niềm vui”. Nếu lập trình viên sử dụng một cụm từ như thế, nó phải được định nghĩa cụ thể. Tuy nhiên, sẽ là không đủ nếu đưa ra định nghĩa cho các khái niệm này bằng các khái niệm cấp cao khác của con người, chẳng hạn như những câu nói triết lý theo kiểu “niềm vui là sự hưởng thụ của các tiềm lực vốn có trong bản chất con người”. Định nghĩa phù hợp phải được đơn giản hóa dưới dạng các thuật ngữ có trong ngôn ngữ lập trình của AI, và cuối cùng phải đưa về được các dữ liệu nguyên thủy như các toán tử và địa chỉ trỏ đến nội dung của từng thanh ghi (memory register) cụ thể. Khi xem xét vấn đề từ góc độ này, chúng ta có thể dần dần hiểu được những khó khăn mà một lập trình viên phải đối mặt khi thực hiện nhiệm vụ.

Xác định và mã hóa những mục tiêu tối thượng của chúng ta là một việc khó khăn, vì các phép biểu diễn mục tiêu của con người rất phức tạp. Sự phức tạp này là rất rõ ràng đối với con người, nhưng chúng ta lại thường không thể công nhận sự hiện diện của nó. Ta có thể so sánh trường hợp này với quá trình tiếp nhận hình ảnh. Nhìn có thể được coi là một việc đơn giản, bởi chúng ta thường làm thế một cách dễ dàng.³ Chỉ cần ta mở mắt, tâm trí sẽ tràn ngập những hình ảnh ba chiều sống động, giàu ý nghĩa và có độ chính xác cao về cảnh vật xung quanh. Hiểu biết vốn có này về thị giác cũng không khác là bao so với việc một vị vua hiểu rõ cung điện của mình^a: chỉ cần ông chú tâm, mọi thứ sẽ xuất hiện đúng nơi, đúng lúc, trong khi đó, những cơ chế đằng sau sự hiện diện này lại khuất mắt trông coi. Tuy nhiên, ngay cả nhiệm vụ quan sát đơn giản nhất, như tìm lọ hạt tiêu trong chạn bếp, cũng đòi hỏi một khối lượng tính toán khổng lồ.

^a Đã chuyển ngữ tương đương cho phù hợp với độc giả Việt Nam. (ND)

Từ một chuỗi thời gian nhiễu loạn của các mẫu hình kích hoạt dây thần kinh hai chiều bắt nguồn từ giác mạc và được truyền lên não thông qua dây thần kinh thị giác, vỏ não thị giác sẽ phải thực hiện công việc theo chiều ngược lại để tạo dựng một phép biểu diễn ba chiều đã được diễn dịch về không gian bên ngoài. Một phần đáng kể trên diện tích bề mặt vỏ não trải rộng một mét vuông của chúng ta được “khoanh vùng” để xử lý thông tin thị giác, và khi bạn đọc cuốn sách này, hàng tỷ neuron đang làm việc không ngừng nghỉ để giúp bạn hoàn thành nhiệm vụ (giống như những thợ may cần mẫn đang gò mình trên máy may để máy đi máy lại một tấm chăn khổng lồ). Theo cách tương tự, các giá trị và ước muốn tưởng chừng đơn giản

của chúng ta vốn bao hàm một sự phức tạp vô biên.⁴ Vậy, làm thế nào mà một lập trình viên có thể truyền tải một hệ thống phức tạp như vậy vào chỉ một hàm thỏa dụng?

Ta có thể thử phương pháp tiếp cận sau đây để mã hóa trực tiếp một phép biểu diễn hoàn chỉnh của bất kỳ mục tiêu nào muốn AI theo đuổi, hay nói cách khác, để viết được một hàm thỏa dụng tường minh. Phương pháp này có thể đạt hiệu quả nếu ta có những mục tiêu tối giản, chẳng hạn như tính toán các chữ số trong số pi, nghĩa là, nếu mục tiêu duy nhất mà ta muốn AI thực hiện chỉ là tính toán các chữ số của số pi và không để tâm đến bất kỳ hệ quả nào khác có thể phát sinh từ việc theo đuổi mục tiêu này (hãy nhớ lại những gì mà chúng ta đã thảo luận về chế độ sai hỏng dẫn đến hiện tượng dư thừa hạ tầng. Phương pháp mã hóa tường minh này có thể mang nhiều hứa hẹn trong quá trình áp dụng phương pháp lựa chọn động lực tự hạn định (domesticity). Nhưng nếu ta tìm cách thúc đẩy hoặc bảo vệ bất kỳ giá trị hợp lý nào của con người, và đang xây dựng một hệ thống nhằm mục đích đạt được siêu trí tuệ tối cao, thì việc mã hóa trực tiếp một phép biểu diễn mục tiêu hoàn chỉnh cần có xem ra sẽ nằm ngoài tầm với.⁵

Chúng ta có thể làm gì nếu không truyền tải được các giá trị của con người vào AI thông qua việc tạo dựng các phép biểu diễn hoàn chỉnh bằng mã máy tính? Chương này sẽ bàn về một số con đường thay thế. Thoạt nhìn, một số con đường trong đó có thể là khả thi, nhưng khi nghiên cứu kỹ hơn, mọi thứ sẽ không còn được như vậy. Công cuộc nghiên cứu khám phá trong tương lai nên tập trung tìm tòi những lối đi còn để ngỏ này.

Vấn đề truyền tải giá trị là một thử thách lớn trong nghiên cứu, đáng để những tài năng toán học xuất chúng của các thế hệ sau đào sâu giải quyết. Con người không thể trì hoãn hành trình đi tìm lời giải cho bài toán này cho đến khi AI đã có đủ lý trí để hiểu được mục đích của chúng ta. Như đã thấy trong phần nội dung bàn về những lý do hội tụ công cụ, một hệ thống đặc trưng sẽ chống lại bất kỳ nỗ lực nào được đưa ra nhằm thay đổi những giá trị tối cao của nó. Một tác tử chưa đạt được mức độ thân thiện cơ bản vào thời điểm nó có được khả năng tự nhận thức về hành động của mình sẽ không tỏ ra nhã nhặn trước một nỗ lực tẩy não muộn màng hay một kế hoạch thay thế nó bằng một tác tử khác thân thiện hơn.

Chọn lọc tiến hóa

Ít nhất một lần, quá trình tiến hóa đã tạo ra một sinh vật sở hữu các giá trị nhân văn. Điều này có thể khích lệ niềm tin rằng các phương pháp tiến hóa chính là cách giải quyết vấn đề truyền tải giá trị. Tuy nhiên, lối đi này cũng ẩn chứa những trở ngại lớn đối với công cuộc tìm kiếm sự an toàn. Chúng ta từng nói về những trở ngại này trong phần cuối của Chương 10 khi bàn về sự nguy hiểm của các quá trình tìm kiếm mạnh mẽ.

Tiến hóa có thể được coi là một loại thuật toán tìm kiếm cụ thể liên quan đến sự xen kẽ của hai thao tác. Một trong số đó giúp gia tăng số lượng các “nghiệm” tiềm năng bằng cách tạo ra nghiệm mới dựa trên một quy tắc ngẫu nhiên tương đối đơn giản (chẳng hạn như đột biến ngẫu nhiên hay tái tổ hợp hữu tính), trong khi đó thao tác còn lại sẽ cắt giảm số lượng bằng cách loại bỏ các nghiệm cho ra kết quả thấp khi được kiểm tra bằng một hàm đánh giá. Giống với các thuật toán tìm kiếm mạnh mẽ khác, ở đây cũng tồn tại một rủi ro: tìm

ra một nghiệm thỏa mãn các tiêu chí tìm kiếm được chỉ định chính thức nhưng lại không thỏa mãn những kỳ vọng hàm định của chúng ta. (Điều này luôn đúng, kể cả khi ta đang tìm cách phát triển một bộ não kỹ thuật số có cùng mục tiêu và giá trị với con người, hay thay vào đó là một bộ não, ví dụ, có đạo đức hoặc có sự phục tùng vô điều kiện.) Ta sẽ tránh được rủi ro này nếu có thể cụ thể hóa một tiêu chí tìm kiếm chính thức biểu thị một cách chính xác toàn bộ các chiều kích của mục tiêu thay vì chỉ một khía cạnh của những gì ta nghĩ là mình muốn. Tuy nhiên, đây lại chính là vấn đề truyền tải giá trị, và đương nhiên, sẽ là nguy hiểm trong hoàn cảnh này nếu ta cho rằng vấn đề đã được giải quyết.

Ngoài ra, ta còn có một vấn đề sâu xa hơn:

Nếu được định lượng mỗi năm, sự đau khổ và chịu đựng trong giới tự nhiên sẽ vượt xa tất cả những gì mà con người nghĩ đến. Vào chính thời điểm tôi viết câu này, hàng nghìn cá thể động vật hoặc bị ăn tươi nuốt sống, hoặc đang cố thoát thân trong sợ hãi. Trong khi đó các cá thể khác lại đang dần bị gặm nhấm từ bên trong bởi đám ký sinh trùng, và vô số sinh vật đang chết dần vì đói khát và bệnh tật.⁶

Ngay chỉ riêng bản thân loài người, mỗi ngày hiện có 150.000 người thiệt mạng cùng vô số người khác đang phải chịu đựng hàng loạt các hành vi ngược đãi và chiếm đoạt.⁷ Thiên nhiên có thể là một nhà thực nghiệm vĩ đại, nhưng lại không bao giờ vượt qua được những cuộc kiểm tra với một danh mục kiểm điểm đạo đức khi ngang nhiên chống lại Tuyên ngôn Helsinki cũng như mọi chuẩn mực về đạo đức. Điều quan trọng nằm ở chỗ chúng ta không được phép tái lập những điều kinh hoàng như vậy khi tạo ra một thế giới trên máy

tính. Xem ra ta khó có thể tránh được những “tội ác tâm trí” khi sử dụng các phương pháp tiến hóa để tạo ra trí tuệ như con người, ít nhất là khi cố tình làm cho quá trình này tương đồng với tiến hóa sinh học.⁸

Học tăng cường

Học tăng cường là một lĩnh vực học máy nghiên cứu các kỹ thuật mà các tác tử có thể học theo để tối đa hóa một ý niệm nào đó về phần thưởng tích lũy (cumulative reward). Bằng cách xây dựng một môi trường mà “phần thưởng” sẽ được trao nếu đạt được hiệu năng mong muốn, một tác tử học tăng cường có thể được tạo ra để học cách giải quyết nhiều vấn đề đa dạng (ngay cả khi các lập trình viên không đưa ra hướng dẫn hoặc phản hồi chi tiết, ngoại trừ dấu hiệu về phần thưởng). Thuật toán học thường bao gồm việc xây dựng dần dần một số loại hàm đánh giá có chức năng gán giá trị cho các trạng thái, các cặp trạng thái-hành động hay các chính sách. (Chẳng hạn, một chương trình có thể học cách chơi trò backgammon thông qua học tăng cường để cải thiện dần việc đánh giá các thế cờ khả thi.) Hàm đánh giá được cập nhật liên tục từ kinh nghiệm có thể được coi là hàm góp phần tạo ra một hình thức học về giá trị. Tuy nhiên, những gì được học không phải là các giá trị mới cuối cùng, mà là các ước tính ngày càng chính xác về giá trị công cụ của việc đạt được các trạng thái cụ thể (hoặc của việc thực hiện các hành động cụ thể ở các trạng thái cụ thể, hay việc tuân thủ các chính sách cụ thể). Tới một chừng mực mà một tác tử học tăng cường có thể được mô tả là có một mục tiêu tối thượng thì mục tiêu đó luôn không đổi, đó chính là nhằm tối đa hóa phần thưởng trong tương lai. Phần thưởng này bao gồm các đối tượng tri giác được chỉ định đặc biệt

mà tác tử nhận được từ môi trường. Vì thế, hội chứng wireheading vẫn là một kết quả khả thể trong bất kỳ tác tử học tăng cường nào phát triển một mô hình thế giới đủ tinh vi để đưa ra cách thức thay thế này cho việc tối ưu hóa phần thưởng.⁹

Những ý kiến này không hàm chỉ rằng các phương pháp học tăng cường không bao giờ có thể được sử dụng trong một AI hạt giống an toàn mà chỉ có hàm ý rằng các phương pháp đó sẽ phải phụ thuộc vào một hệ thống động lực không tự được tổ chức xoay quanh nguyên tắc tối ưu hóa phần thưởng. Tuy nhiên, điều này sẽ đòi hỏi một số phương pháp khác ngoài học tăng cường tìm ra trước một giải pháp cho vấn đề truyền tải giá trị.

Bồi đắp giá trị liên kết (associative value accretion)

Bây giờ có thể ai đó sẽ tự hỏi: nếu vấn đề truyền tải dữ liệu quá khó thì làm thế nào chúng ta có thể tự thụ đắc các giá trị của mình?

Một mô hình khả thể (bị tối giản hóa quá mức) có thể sẽ như sau. Chúng ta bắt đầu cuộc sống của mình với một vài sở thích hay sở ghét tương đối đơn giản (ví dụ như sự ác cảm với các yếu tố kích thích độc hại), cùng với một tập hợp các khuynh hướng thụ đắc sở thích và sở ghét bổ sung như một sự phản hồi với các trải nghiệm khả thể (ví dụ: chúng ta có thể có khuynh hướng tạo ra một sự yêu thích đối với các đối tượng và hành vi được coi là có giá trị và đáng tưởng thưởng trong nền văn hóa của mình). Cả các sở thích/sở ghét đơn giản và các khuynh hướng đều là bẩm sinh, và đã được định hình bởi quá trình chọn lọc tự nhiên và giới tính trong suốt thời gian tiến hóa. Tuy nhiên, những sở thích/sở ghét mà chúng ta có được khi trưởng thành lại phụ thuộc vào các sự kiện xảy ra trong đời. Vì thế,

phần lớn nội dung thông tin trong các giá trị cuối cùng của chúng ta có được từ kinh nghiệm chứ không được “tải” sẵn vào bộ gen.

Chẳng hạn, đa số chúng ta đều có người mà mình yêu thương, và sẽ đặt cho sự bình an của họ một giá trị cuối cùng to lớn. Vậy để biểu diễn một giá trị như thế đòi hỏi những gì? Việc này liên quan đến rất nhiều yếu tố, nhưng hãy xem xét hai yếu tố sau: một phép biểu diễn cho “người” và một phép biểu diễn cho “sự bình an”. Các khái niệm này không được mã hóa trực tiếp trong ADN. Thay vào đó, ADN của chúng ta chứa các chỉ lệnh về việc xây dựng một bộ não mà khi được đặt vào một môi trường cơ thể người đặc trưng, sau vài năm sẽ phát triển một mô hình thế giới có chứa khái niệm về mọi người và sự bình an. Khi đã hình thành, các khái niệm này có thể được dùng để biểu diễn một số giá trị có ý nghĩa nhất định. Nhưng, chúng ta cũng cần có một cơ cấu nào đó từ khi chào đời để dẫn đến các giá trị được hình thành quanh chính các khái niệm này chứ không phải bất kỳ khái niệm nào khác có được (như lọ hoa hay dụng cụ mở nắp chai).

Những chi tiết về cách hoạt động của cơ chế này vẫn chưa được thực sự hiểu rõ, và trong cơ thể con người, cơ chế này có thể phức tạp và đa dạng. Hiện tượng này sẽ dễ hiểu hơn nếu chúng ta xem xét nó dưới một dạng thức “thô sơ” hơn, chẳng hạn như hiện tượng “in dấu mẫu tử” (filial imprinting) của những loài chim có thể sống độc lập khi mới ra đời. Trong các loài này, con non ham muốn được ở gần một đối tượng có những yếu tố kích thích chuyển động phù hợp trong ngày đầu tiên sau khi nở. Việc con non muốn ở bên đối tượng cụ thể nào phụ thuộc vào trải nghiệm của bản thân nó, còn di truyền chỉ xác định được thiên hướng in dấu (imprinting) theo cách này mà

thôi. Tương tự như vậy, một người, tên là Harry chẳng hạn, có thể đặt một giá trị cuối cùng lên sự bình an của người mình yêu thương, giả sử là Sally. Tuy nhiên, nếu cặp đôi này không bao giờ gặp nhau, Harry có thể yêu một người khác, và những giá trị cuối cùng của anh ta có thể sẽ khác đi. Khả năng mã hóa để xây dựng một cơ chế theo đuổi mục tiêu của các gen giải thích cách mà chúng ta có được những mục đích cuối cùng với thông tin vô cùng phức tạp, hơn mức có thể hàm chứa được trong bản thân bộ gen.

Vì thế, chúng ta có thể xem xét xem liệu mình có xây dựng được hệ thống động lực cho một trí tuệ nhận tạo theo cùng nguyên tắc hay không. Nghĩa là, thay vì trực tiếp xác định các giá trị phức tạp, liệu chúng ta có thể xác định một cơ chế nào đó dẫn tới việc thụ đắc những giá trị này khi AI tương tác với một môi trường phù hợp?

Việc sao chép quá trình bồi đắp giá trị ở con người xem ra không dễ dàng. Cơ chế di truyền có liên quan ở người là sản phẩm được tạo ra sau một thời gian làm việc tưởng như vô tận của quá trình tiến hóa mà ta khó có thể tóm lược. Hơn nữa, cơ chế này có thể đã được “đo ni đóng giày” cho kiến trúc thần kinh nhận thức của con người, và vì thế, không áp dụng được trong các dạng trí tuệ máy nếu không phải là giả lập hoàn chỉnh não bộ. Và nếu chúng ta đã sẵn có các dạng giả lập hoàn chỉnh não bộ với mức độ chân thực phù hợp, xem ra mọi việc sẽ dễ dàng hơn nếu bắt đầu từ một bộ não người lớn đã được tải sẵn những phép biểu diễn đầy đủ của một số giá trị nhân bản.¹⁰

Vì thế, việc tìm cách thực thi một quá trình bồi đắp giá trị giống hệt quá trình sinh học ở người xem ra không phải là một hướng tấn công hứa hẹn nhắm tới vấn đề truyền tải giá trị. Nhưng hẳn là chúng

ta sẽ thiết kế được một cơ chế thay thế nhân tạo mạnh mẽ hơn, dẫn đến việc một AI có thể nhập các phép biểu diễn có độ chân thực cao của các giá trị phức tạp có liên quan vào hệ thống mục tiêu của nó. Để việc này thành công, việc trao cho AI những thiên hướng đánh giá giống hệt như của con người. Ngay cả khi là một mục tiêu thì việc này có thể cũng không mấy hấp dẫn – bản chất con người rất cuộc cũng đầy khiếm khuyết và thường xuyên hé lộ khuynh hướng thiên về cái ác. Điều này là không thể chấp nhận được trong bất cứ hệ thống nào đã sẵn sàng có được một ưu thế chiến lược quyết định. Tốt hơn, có lẽ ta nên nhắm tới một hệ thống động lực rời khỏi các chuẩn mực của con người một cách có hệ thống, chẳng hạn như có một xu hướng mạnh mẽ hơn về những mục tiêu tối thượng mang tính vị tha, từ bi hoặc cao cả, theo những cách có thể được công nhận là phản ánh nhân cách vô cùng tốt đẹp nếu xuất hiện ở con người. Tuy nhiên, để được coi là cải tiến, những sai khác này so với chuẩn mực của con người phải được định hướng rất cụ thể thay vì ngẫu nhiên; và chúng sẽ tiếp tục giả định trước sự tồn tại của một bộ khung tham chiếu mang tính dĩ nhân vi trung đa phần không bị nhiễu loạn để mang lại những dạng khái quát hóa mang tính đánh giá có ý nghĩa nhân bản (nhằm tránh được việc cụ thể hóa không đúng cách những mô tả về mục tiêu khả thi nhưng hời hợt mà chúng ta đã xem xét trong Chương 8). Tính khả thi của điều này vẫn còn là một câu hỏi mở.

Một vấn đề nữa của việc bồi đắp giá trị liên kết là AI có thể vô hiệu hóa cơ chế bồi đắp. Như đã thấy trong Chương 7, tính nhất quán mục tiêu-hệ thống là một giá trị hội tụ công cụ. Khi AI đạt đến một giai đoạn phát triển nhận thức nhất định, nó có thể bắt đầu xem

việc liên tục vận hành cơ chế bồi đắp này như một ảnh hưởng gây lỗi.¹¹

Đây không hẳn là một điều xấu, nhưng chúng ta cần chú ý sao cho việc “niêm phong” hệ thống mục tiêu diễn ra đúng thời điểm, sau khi các giá trị thích hợp đã được bồi đắp, và trước khi chúng bị ghi đè bởi các giá trị bồi đắp bổ sung không mong muốn.

Khung tạo động lực

Một cách tiếp cận khác đến vấn đề truyền tải giá trị là khung tạo động lực (motivational scaffolding). Khung này liên quan đến việc cấp cho AI hạt giống một hệ thống mục tiêu tạm thời, với những mục tiêu tối thượng tương đối đơn giản có thể biểu diễn bằng cách mã hóa tường minh (explicit coding) hoặc một phương pháp khả thi khác. Một khi AI đã phát triển được những năng lực đại diện phức tạp hơn, chúng ta sẽ thay thế hệ thống mục tiêu khung tạm thời này bằng một hệ thống có các mục tiêu tối thượng khác. Hệ thống mục tiêu kế tục này sau đó sẽ kiểm soát AI khi nó phát triển thành một siêu trí tuệ toàn diện.

Vì các mục tiêu khung không chỉ mang tính công cụ mà là các mục tiêu tối thượng, ta có thể dự đoán rằng AI sẽ kháng cự việc thay thế chúng (tính nhất quán mục tiêu-nội dung là một giá trị hội tụ công cụ). Việc này tạo ra một mối nguy hại. Nếu AI cản trở thành công việc thay thế các mục tiêu khung của nó thì phương pháp này sẽ thất bại.

Các biện pháp phòng ngừa cần được áp dụng để tránh việc thất bại như vậy. Ví dụ, các phương pháp kiểm soát năng lực có thể được dùng để hạn chế năng lực của AI cho tới khi hệ thống động lực hoàn thiện được cài đặt. Cụ thể, ta có thể tìm cách hạn chế sự phát triển về nhận thức của AI ở một mức độ an toàn, nhưng vẫn cho

phép nó biểu diễn các giá trị mà chúng ta muốn đưa vào trong các mục tiêu tối thượng. Để làm được việc này, chúng ta có thể dùng nhiều cách để kìm hãm một số loại năng lực trí tuệ nhất định, như những năng lực cần để lập chiến lược và bày mưu tính kế, trong khi cho phép những năng lực (tỏ ra) vô hại hơn được phát triển tới một cấp độ cao hơn.

Chúng ta cũng có thể sử dụng các phương pháp lựa chọn động lực để đem đến một mối quan hệ hợp tác sâu hơn giữa AI hạt giống và nhóm lập trình viên. Chẳng hạn, ta có thể đưa vào hệ thống động lực khung một mục tiêu cởi mở với những hướng dẫn trực tuyến từ các lập trình viên, bao gồm cả việc cho phép họ thay thế bất cứ mục tiêu hiện tại nào của AI.¹² Những mục tiêu khung khác có thể bao gồm việc bộc lộ rõ ràng với lập trình viên các giá trị và chiến lược của AI, đồng thời phát triển một kiến trúc dễ hiểu đối với lập trình viên và tạo điều kiện cho việc triển khai sau đó của một mục tiêu tối thượng có ý nghĩa với con người, cũng như các động lực tự hạn định (như hạn chế việc sử dụng các nguồn lực tính toán).

Ai đó thậm chí cũng có thể hình dung ra việc cấp cho AI hạt giống mục tiêu tối thượng duy nhất là thay thế nó bằng một mục tiêu tối thượng khác, có thể chỉ được lập trình viên xác định một cách ẩn ý hay gián tiếp. Một số vấn đề phát sinh từ việc sử dụng mục tiêu khung “tự thay thế” này cũng có thể nảy sinh trong bối cảnh của phương pháp học giá trị sẽ được thảo luận trong phần sau. Chúng ta cũng sẽ thảo luận một số vấn đề sâu hơn trong Chương 13.

Cách tiếp cận khung tạo động lực cũng có một số nhược điểm. Trong số đó, cách tiếp cận này mang theo rủi ro rằng AI có thể trở nên quá mạnh mẽ khi vẫn còn đang vận hành trong hệ thống mục

tiêu tạm thời của mình. Sau đó, nó có thể ngăn cản các nỗ lực của lập trình viên trong việc cài đặt hệ thống mục tiêu tối thượng (bằng cách kháng cự mạnh mẽ hoặc âm thầm lật đổ). Các mục tiêu tối thượng đã cũ sau đó vẫn có thể hoạt động khi AI hạt giống phát triển thành một siêu trí tuệ toàn diện. Nhược điểm khác nằm ở chỗ việc cài đặt các mục tiêu dự kiến cuối cùng vào một AI cấp độ con người không hẳn đã dễ dàng hơn đến vậy so với việc làm điều tương tự trên một AI sơ khai hơn. Một AI cấp độ con người phức tạp hơn và có thể đã phát triển một kiến trúc không rõ ràng và khó thay đổi. Một AI hạt giống, ngược lại, lại giống như một tờ giấy trắng, trên đó các lập trình viên có thể viết bất cứ cấu trúc nào mà họ coi là hữu ích. Nhược điểm này có thể biến thành ưu điểm nếu chúng ta thành công trong việc cho AI hạt giống những mục tiêu khung, khiến nó muốn phát triển một kiến trúc hữu ích đối với các lập trình viên khi họ nỗ lực cài đặt các giá trị cuối cùng sau này. Tuy nhiên, ta vẫn chưa biết rõ liệu việc cấp cho AI hạt giống những mục tiêu khung với tính chất như vậy sẽ dễ dàng đến đâu, cũng như làm thế nào để ngay cả một AI hạt giống đã được tạo động lực một cách lý tưởng có được khả năng phát triển một cấu trúc tốt với chất lượng cao hơn rất nhiều so với nhóm lập trình viên.

Học giá trị

Bây giờ chúng ta sẽ bàn về một phương pháp tiếp cận quan trọng và tinh tế với vấn đề truyền tải giá trị. Phương pháp này liên quan đến việc sử dụng trí tuệ của AI để học các giá trị mà chúng ta muốn nó theo đuổi. Để làm được việc này, chúng ta phải cung cấp một tiêu chí sao cho AI chọn ra được một tập giá trị thích hợp nào đó, chí ít là một cách hàm ẩn. Sau đó, chúng ta có thể tiến hành xây

dựng để AI hành động theo những ước tính tốt nhất của nó về những giá trị được định nghĩa theo cách hàm ẩn này. Nó sẽ liên tục hoàn thiện những ước tính khi học hỏi được nhiều hơn về thế giới, và dần tháo gỡ được những hàm ý của tiêu chí xác định giá trị.

Trái với khung tạo động lực (cho AI một mục tiêu khung tạm thời và sau đó thay nó bằng một mục tiêu tối thượng khác), phương pháp học giá trị duy trì một mục tiêu tối thượng không đổi trong suốt các giai đoạn phát triển và vận hành của AI. Việc học tập không làm thay đổi mục tiêu tối thượng mà chỉ thay đổi niềm tin của AI về mục tiêu đó.

Vì thế, AI phải được trang bị một tiêu chí mà nó có thể sử dụng để xác định các đối tượng tri giác tạo thành bằng chứng cho một giả thuyết nào đó về mục tiêu tối thượng, đồng thời xác định các đối tượng tri giác tạo thành bằng chứng chống lại giả thuyết đó. Xác định một tiêu chí phù hợp có thể là khá khó khăn, nhưng một phần của khó khăn này lại liên quan đến vấn đề tạo ra trí tuệ nhân tạo tổng thể ngay từ ban đầu. Vấn đề này đòi hỏi một cơ chế học tập mạnh mẽ có thể khám phá được cấu trúc của môi trường từ những dữ liệu giác quan đầu vào hạn chế. Bây giờ chúng ta có thể tạm gác vấn đề này sang một bên. Nhưng ngay cả khi đã có một giải pháp cho vấn đề làm thế nào để tạo ra một AI siêu trí tuệ thì vẫn còn lại những khó khăn đặc biệt phát sinh từ vấn đề truyền tải giá trị. Với phương pháp học giá trị, những khó khăn này sẽ hiện diện dưới dạng nhu cầu đối với việc xác định một tiêu chí liên kết các chuỗi bit (bitstring) với các giả thuyết về giá trị.

Trước khi đào sâu vào các chi tiết của việc làm thế nào để triển khai phương pháp học giá trị, một ví dụ có thể sẽ hữu ích trong việc

minh họa ý tưởng chung. Giả sử chúng ta viết lên giấy một đoạn mô tả về một tập hợp các giá trị, gấp tờ giấy lại cho vào phong bì, sau đó tạo ra một tác tử với trí tuệ tổng quát cấp độ con người và cho nó mục tiêu tối thượng sau đây: “Hãy tối đa hóa việc hiện thực hóa những giá trị được mô tả trong phong bì.” Tác tử này sẽ làm gì?

Ban đầu, tác tử không biết trong phong bì viết gì. Tuy nhiên, nó có thể đặt ra các giả thiết, và có thể ấn định cho những giả thiết này các xác suất dựa trên các tiền lệ và bất cứ dữ liệu thực nghiệm sẵn có nào. Tác tử có thể gặp phải các văn bản khác do con người viết ra, hoặc có thể quan sát được một số mẫu hình khái quát của hành vi con người. Điều này cho nó khả năng tiên đoán. Chúng ta không cần một tấm bảng tâm lý học để dự báo rằng tờ giấy nhiều khả năng sẽ mô tả những giá trị như “giảm thiểu bất công và đau khổ không cần thiết” hoặc “tối đa hóa hoạt động hoàn vốn cho cổ đông” hơn là những giá trị như “lắp tất cả hồ ao bằng túi nylon”.

Khi một tác tử ra quyết định, nó tìm cách thực hiện những hành động có hiệu quả trong việc hiện thực hóa các giá trị mà nó tin là có nhiều khả năng được mô tả trong bức thư nhất. Quan trọng hơn, tác tử có thể nhìn thấy một giá trị công cụ cao trong việc tìm hiểu thêm về ý nghĩa của bức thư. Đó là vì gần như mọi giá trị cuối cùng có thể được mô tả trong bức thư đều có khả năng được hiện thực hóa cao hơn nếu tác tử biết được nó là gì, vì tác tử sau đó sẽ theo đuổi giá trị này một cách có hiệu quả. Tác tử cũng sẽ khám phá những lý do hội tụ công cụ đã được mô tả trong Chương 7 – tính toàn vẹn mục tiêu-hệ thống, tăng cường nhận thức, thụ đắc tài nguyên,... Mặc dù vậy, nếu tác tử gán một xác suất đủ cao cho các giá trị mô tả trong bức thư có liên quan đến lợi ích của con người, nó sẽ không theo đuổi

những giá trị công cụ này bằng cách lập tức biến Trái đất thành computronium rồi tiêu diệt loài người, vì làm như vậy sẽ gây ra nguy cơ triệt tiêu vĩnh viễn khả năng hiện thực hóa giá trị cuối cùng của nó.

Chúng ta có thể so sánh loại tác tử này với một sà lan được gắn với vài chiếc tàu kéo và mỗi chiếc chạy theo một hướng, trong đó, mỗi tàu kéo tương ứng với một giả thiết về giá trị cuối cùng của tác tử. Công suất động cơ của mỗi tàu kéo tương ứng với xác suất của giả thiết liên quan, và vì vậy những thay đổi, chẳng hạn như khi có bằng chứng mới, sẽ điều chỉnh hướng chuyển động của sà lan. Lực được tạo ra phải kéo được chiếc xà lan theo một quỹ đạo tạo điều kiện cho việc học tập về giá trị cuối cùng (dạng hàm ẩn), đồng thời tránh được bãi cạn dẫn đến việc bị triệt tiêu không thể hồi phục; và sau đó, khi tới được cửa biển của những kiến thức cụ thể hơn của giá trị cuối cùng, chiếc tàu kéo duy trì được sức kéo đáng kể sẽ kéo sà lan về hướng hiện thực hóa giá trị đã khám phá theo tuyến đường thẳng nhất hoặc thuận lợi nhất. Ẩn dụ về cái phong bì và chiếc xà lan minh họa cho nguyên tắc của phương pháp học giá trị, nhưng lại bỏ qua một số vấn đề kỹ thuật quan trọng. Chúng sẽ tiến vào vùng tiêu điểm rõ ràng hơn khi chúng ta bắt đầu phát triển cách tiếp cận bên trong một khuôn khổ mang tính hình thức (xem Khung 10).

Một vấn đề nổi cộm khác là làm thế nào để gán cho AI một mục tiêu như: “Tối đa hóa việc hiện thực hóa các giá trị được mô tả trong phong bì.” (Theo thuật ngữ của Khung 10, là làm thế nào để xác định tiêu chí giá trị V). Để làm việc này, cần xác định vị trí mà ở đó các giá trị được mô tả. Trong ví dụ của chúng ta, điều này đòi hỏi thực hiện một tham chiếu thành công tới bức thư trong phong bì. Việc này tuy

có vẻ bình thường, nhưng không phải không có cạm bẫy. Một trong số đó là: điều rất quan trọng là không chỉ đơn giản được tham chiếu tới một đối tượng vật lý cụ thể bên ngoài mà tới một đối tượng tại một thời điểm nhất định. Nếu không, AI có thể xác định rằng cách tốt nhất để đạt được mục tiêu của nó là ghi đè lên phần mô tả giá trị gốc một giá trị khác mang lại mục tiêu dễ dàng hơn (chẳng hạn như giá trị quy định với mỗi số nguyên sẽ có một số nguyên lớn hơn). Khi hoàn thành việc này, AI có thể ngả người bẻ đốt ngón tay, dù một thất bại nghiêm trọng khả năng cao là sẽ xảy ra vì những lý do đã bàn đến trong Chương 8.

Bây giờ chúng ta sẽ đối mặt với câu hỏi về việc xác định thời gian. Chúng ta có thể chỉ vào đồng hồ và nói, “Thời gian được xác định thông qua chuyển động của thiết bị này”, nhưng điều này có thể sai nếu AI phỏng đoán rằng nó có thể thao túng thời gian bằng cách di chuyển kim đồng hồ, và phỏng đoán này là hoàn toàn đúng khi “thời gian” được định nghĩa như vậy. (Trong một trường hợp thực tế, vấn đề có thể trở nên phức tạp hơn vì những giá trị liên quan có thể không được mô tả một cách rõ ràng trong bức thư; và nhiều khả năng, chúng sẽ phải được suy luận từ việc quan sát các cấu trúc đã tồn tại có chứa thông tin liên quan một cách hàm ẩn, chẳng hạn như não người.)

Khung 10: Hình thức hóa phương pháp học giá trị

Khung này giới thiệu một khái niệm mang tính hình thức có thể giúp chúng ta nhìn sự vật một cách rõ ràng hơn. Tuy nhiên các độc giả không thích chủ nghĩa hình thức có thể bỏ qua phần này.

Hãy xem xét một khung đã được đơn giản hóa, trong đó một tác tử tương tác với môi trường của nó trong một số chu kỳ rời rạc.¹³

Trong chu kỳ k , tác tử thực hiện hành động y_k rồi nhận được đối tượng tri giác x_k . Lịch sử tương tác của một tác tử với vòng đời m là một chuỗi $y_1x_1y_2x_2\ldots y_mx_m$ (ta có thể viết tắt chuỗi này là $yx_{1:m}$ hoặc $yx_{\leq m}$). Trong mỗi chu kỳ, tác tử chọn một hành động dựa trên chuỗi đối tượng tri giác mà nó đã nhận được.

Trước hết hãy xem xét một đối tượng học tăng cường. Một đối tượng học tăng cường tối ưu (AI-RL) là đối tượng tối ưu hóa được các phần thưởng mong đợi trong tương lai. Nó tuân theo phương trình:¹⁴

$$y_k = \arg \max_{y_k} \sum_{x_k y_{k+1:m}} (r_k + \dots + r_m) P(yx_{\leq m} | yx_{< k} y_k).$$

Chuỗi phần thưởng $r_k, \dots, r_m$ được bao hàm bởi chuỗi đối tượng tri giác x_{km} , vì phần thưởng mà tác tử nhận được trong một chu kỳ nhất định là một phần của đối tượng tri giác mà tác tử nhận được trong chu kỳ đó.

Như đã được thảo luận ở trên, loại hình học tăng cường này không thích hợp trong bối cảnh hiện tại vì một tác tử đủ thông minh sẽ nhận ra rằng nó có thể đảm bảo có được phần thưởng tối đa nếu trực tiếp điều khiển tín hiệu phần thưởng của nó (wireheading). Đối với những tác tử yếu hơn, điều này không hẳn là vấn đề, bởi lẽ ta có thể ngăn không cho chúng tác động đến kênh phần thưởng của chính mình bằng cách thức vật lý. Chúng ta cũng có thể kiểm soát môi trường của chúng sao cho chúng chỉ nhận được phần thưởng khi hành xử theo cách mà chúng ta chấp nhận. Tuy nhiên, một đối

tượng học tăng cường như vậy lại có động cơ rất mạnh trong việc loại bỏ sự phụ thuộc nhân tạo này của phần thưởng vào mong muốn và ý tưởng của chúng ta. Mỗi quan hệ của chúng ta với một đối tượng học tăng cường vì thế mang tính đối lập cơ bản. Nếu một tác tử trở nên mạnh mẽ, điều này cũng ẩn chứa nguy hiểm.

Các biến thể của hội chứng wireheading có thể còn ảnh hưởng tới các hệ thống không tìm kiếm tín hiệu giác quan về phần thưởng từ bên ngoài, mà có mục tiêu được xác định là việc đạt đến một trạng thái bên trong nào đó.

Ví dụ, trong các hệ thống được gọi là “tác thể-phê phán” (actor-critic), module tác thể (actor module) sẽ lựa chọn các hành động để tối thiểu hóa sự phản đối của module phê phán (critic module) riêng biệt, một module có nhiệm vụ tính toán xem hành vi của tác tử còn cách biệt bao xa so với một giá trị đo hiệu năng cho trước. Vấn đề của thiết lập này nằm ở chỗ module tác thể có thể nhận ra rằng nó có khả năng giảm thiểu được sự phê phán bằng cách điều chỉnh hoặc thậm chí loại bỏ hoàn toàn module phê phán. Điều này rất giống với việc một nhà độc tài giải tán quốc hội và quốc hữu hóa báo chí. Đối với các hệ thống hạn chế, vấn đề có thể được tránh khỏi một cách đơn giản thông qua việc không cho module tác thể bất cứ phương tiện gì để sửa đổi module phê phán. Tuy nhiên, một module tác thể đủ thông minh và tháo vát luôn có thể tiếp cận được module phê phán (mà sau tất cả, chỉ là một quá trình vật lý trong máy tính).¹⁵

Trước khi bàn về đối tượng học giá trị, chúng ta hãy xem xét công thức tối ưu hóa quan sát-thỏa dụng (AI-OUM) với tư cách là một bước trung gian.

Công thức này có được nhờ việc thay thế chuỗi phần thưởng ($r_k + \dots + r_m$) trong AI-RL bằng một hàm thỏa dụng cho phép phụ thuộc vào toàn bộ lịch sử tương tác trong tương lai của AI:

$$y_k = \arg \max_{y_k} \sum_{x_k y_{k+1:m}} U(y x_{\leq m}) P(y x_{\leq m} | y x_{< k} y_k).$$

Công thức này cho ta cách giải quyết vấn đề wireheading, vì một hàm thỏa dụng được xác định trên toàn bộ lịch sử tương tác có thể được thiết kế để áp dụng hình phạt với các mục trong lịch sử tương tác cho thấy dấu hiệu của sự tự lừa dối (hoặc của thất bại của tác tử trong việc đầu tư đầy đủ nguồn lực nhằm có được cái nhìn đúng về hiện thực).

AI-OUM vì vậy sẽ khiến cho việc ngăn chặn vấn đề wireheading trở nên khả thi về mặt nguyên tắc. Tuy nhiên, để có được khả năng này, chúng ta phải xác định một hàm thỏa dụng phù hợp trên các mục khả thể trong lịch sử tương tác. Xem ra đây là một nhiệm vụ vô cùng khó khăn.

Có thể sẽ tự nhiên hơn khi xác định các hàm thỏa dụng trực tiếp từ phương diện các thế giới khả thể (hoặc các tính chất của các thế giới khả thể, hay các lý thuyết về thế giới đó) thay vì từ phương diện các mục trong lịch sử tương tác của bản thân tác tử. Nếu sử dụng cách tiếp cận này, chúng ta có thể lập lại công thức và đơn giản hóa khái niệm về tính tối ưu của AI-OUM:

$$y = \arg \max_y \sum_w U(w)P(w|Ey).$$

Tại đây, E là tổng số bằng chứng mà tác tử có được (khi ra quyết định), và U là một hàm thỏa dụng gán độ thỏa dụng cho một lớp nào đó của các thế giới khả thể. Một tác tử tối ưu lựa chọn hành động có thể tối đa hóa độ thỏa dụng kỳ vọng.

Vấn đề còn tồn tại với những phát biểu này là sự khó khăn trong việc xác định hàm thỏa dụng U . Rốt cuộc, điều này lại đưa chúng ta về với vấn đề truyền tải giá trị. Để khởi động hàm thỏa dụng sẽ được học, chúng ta phải mở rộng chủ nghĩa hình thức của mình để cho phép chấp nhận sự bất định trên các hàm thỏa dụng. Điều này có thể được thực hiện như sau (AI-VL):¹⁶

$$y = \arg \max_{y \in Y} \sum_{w \in W} P(w|Ey) \sum_{u \in U} U(w)P(\mathcal{V}(U)|w).$$

Ở đây $\mathcal{V}(\cdot)$ là một hàm từ các hàm thỏa dụng tới các mệnh đề về các hàm thỏa dụng. $\mathcal{V}(U)$ là mệnh đề nói rằng hàm thỏa dụng U thỏa mãn điều kiện giá trị được $\mathcal{V}$ biểu diễn.¹⁷

Để quyết định hành động sẽ thực thi, ta có thể làm theo các bước sau: Thứ nhất, tính toán xác suất có điều kiện của từng thế giới khả thể w (khi bằng chứng là sẵn có và giả định rằng hành động y sẽ được thực hiện). Thứ hai, đối với mỗi hàm thỏa dụng khả thể U , tính

toán xác suất có điều kiện của việc U thỏa mãn tiêu chí giá trị V (có điều kiện w là thế giới thực.) Thứ ba, đối với mỗi hàm thỏa dụng khả thể U, tính toán độ thỏa dụng của thế giới khả thể w. Thứ tư, kết hợp những giá trị này để tính toán độ thỏa dụng kỳ vọng của hành động y. Thứ năm, lặp lại quy trình này cho mỗi hành động khả thể, và thực hiện hành động có độ thỏa dụng kỳ vọng cao nhất (sử dụng một phương pháp tùy ý nào đó để kết thúc quy trình). Như đã mô tả, quy trình (liên quan đến việc đưa ra xem xét cụ thể và riêng biệt cho mỗi thế giới khả thể) này tất nhiên là rất khó khăn về mặt tính toán. AI có thể sử dụng những cách tính toán vắn tắt để tính gần đúng khái niệm về độ tối ưu này.

Câu hỏi đặt ra khi đó sẽ là cách thức xác định tiêu chí giá trị V.¹⁸ Một khi AI có một phép biểu diễn đầy đủ của tiêu chí giá trị, thì về nguyên tắc, nó có thể sử dụng trí tuệ tổng quát của mình để thu thập thông tin về những thế giới khả thể nào có các xác suất là thế giới thực cao nhất.

Sau đó, nó có thể áp dụng tiêu chí này cho từng thế giới khả thể phù hợp w để tìm xem hàm thỏa dụng nào thỏa mãn được tiêu chí V trong w. Do đó, ta có thể coi công thức AI-VL là một cách để xác định và tách biệt vấn đề thách thức then chốt này trong phương pháp học giá trị, là thách thức của việc làm thế nào để biểu diễn tiêu chí V. Chủ nghĩa hình thức cũng giúp làm sáng tỏ một số vấn đề khác (chẳng hạn như làm thế nào để xác định Y, W và U) cần được giải quyết trước khi phương pháp tiếp cận này có được hiệu quả.¹⁹

Một vấn đề khác trong việc mã hóa mục tiêu “tối đa hóa việc hiện thực hóa những giá trị được mô tả trong phong bì” là: Ngay cả nếu toàn bộ các giá trị đúng đã được mô tả trong một bức thư, và hệ

thống động lực của AI đã được chỉ dẫn thành công theo nguồn này, thì AI cũng có thể không diễn giải các mô tả theo cách mà chúng ta mong muốn. Điều này có thể tạo ra rủi ro gặp phải hiện tượng hiện thực hóa sai lệch được thảo luận trong Chương 8.

Nói một cách rõ ràng, khó khăn mà ta gặp phải ở đây không hẳn là việc làm thế nào để đảm bảo rằng AI có thể hiểu được chủ ý của con người. Một siêu trí tuệ sẽ dễ dàng phát triển được sự hiểu biết như thế. Khó khăn thật ra nằm ở việc đảm bảo rằng AI sẽ được tạo động lực để theo đuổi các giá trị được mô tả theo cách mà chúng ta mong muốn. Khả năng hiểu được chủ ý con người của AI không đảm bảo điều này: một AI có thể biết chính xác ý định của chúng ta là gì nhưng vẫn thờ ơ với cách giải thích đó. (Thay vào đó, nó được tạo động lực bằng một cách giải thích khác hoàn toàn với những gì mà chúng ta đã cung cấp.)

Khó khăn này sẽ trở nên phức tạp hơn bởi một ước nguyện rằng, vì lý do an toàn, động lực đúng đắn cần được thiết lập vào AI hạt giống trước khi nó có được đầy đủ khả năng biểu diễn các khái niệm hoặc hiểu được các chủ ý của con người. Điều này đòi hỏi phải tìm ra cách tạo ra một khung nhận thức, và một vị trí cụ thể trong bộ khung đó được chỉ định trong hệ thống động lực của AI như là nơi lưu trữ giá trị cuối cùng của nó. Nhưng bản thân bộ khung lưu trữ cũng phải chỉnh sửa được, nhằm cho phép AI tăng cường năng lực biểu diễn của mình khi nó học hỏi thêm về thế giới và trở nên thông minh hơn. AI có thể trải qua một sự kiện tương đương với các cuộc cách mạng khoa học, trong đó thế giới quan của nó bị chấn động và bản thân nó có thể rơi vào những cuộc khủng hoảng bản thể học mà trong đó, nó phát hiện ra rằng lối tư duy trước đây của mình về các

giá trị đã dựa trên những nhầm lẫn và ảo tưởng. Tuy khởi đầu ở một mức độ phát triển thấp hơn loài người và tiếp tục trải qua toàn bộ quá trình phát triển sau đó để trở thành một siêu trí tuệ hùng mạnh, nhưng hành vi của AI vẫn cần được hướng dẫn bởi một giá trị cuối cùng không đổi, một giá trị được AI hiểu ngày càng rõ hơn như hệ quả trực tiếp của sự tiến bộ về trí tuệ tổng quát, và có thể được AI trưởng thành (mature AI) hiểu theo một cách hoàn toàn khác so với các lập trình viên ban đầu, tuy sự khác biệt này không hề thù địch hay ngẫu nhiên mà hoàn toàn phù hợp. Cách thức thực hiện điều này vẫn còn đang bỏ ngỏ.²⁰ (Xem Khung 11.)

Tóm lại, chúng ta vẫn chưa biết rõ cách thức áp dụng phương pháp học giá trị để thiết lập các giá trị chấp nhận được của con người. (Tuy nhiên, Khung 12 có thể cung cấp một số ví dụ về những ý tưởng hiện nay.) Hiện tại, phương pháp tiếp cận này có thể được xem như một chương trình nghiên cứu chứ không phải một kỹ thuật sẵn có. Nếu hoạt động được, nó có thể trở thành giải pháp lý tưởng nhất cho vấn đề truyền tải giá trị. Ngoài những lợi ích khác, xem ra phương pháp này sẽ đề xuất một cách thức tự nhiên để ngăn chặn tội ác tâm trí, vì AI hạt giống, với những đoán định hợp lý về những giá trị do lập trình viên cài đặt, sẽ phán đoán được rằng tội ác tâm trí có thể sẽ bị đánh giá một cách tiêu cực bởi các giá trị này, và vì thế tốt hơn hết nên được ngăn chặn, ít nhất là cho đến khi có được thêm thông tin.

Cuối cùng, nhưng không kém phần quan trọng là câu hỏi “viết gì vào trong phong bì” – hoặc tường minh hơn, là câu hỏi về việc chúng ta nên cố gắng cho AI học những giá trị nào. Tuy nhiên, đây là câu hỏi chung của toàn bộ các hướng tiếp cận đối với vấn đề truyền tải

giá trị của AI. Chúng ta sẽ quay trở lại với câu hỏi này trong Chương 13.

Khung 11: Một AI muốn trở nên thân thiện

Eliezer Yudkowsky đã cố gắng mô tả một số tính năng của kiến trúc AI hạt giống được mong đợi sẽ khởi động loại hành vi đã mô tả ở trên. Theo thuật ngữ của ông, AI sẽ sử dụng “ngữ nghĩa ngoại tham chiếu” (external reference semantics).²¹ Để minh họa cho ý tưởng cơ bản, hãy giả sử rằng chúng ta muốn hệ thống trở nên “thân thiện”. Hệ thống khởi đầu với mục tiêu cố gắng hiện thực hóa thuộc tính F nhưng ban đầu lại không biết nhiều về F. Có thể nó chỉ biết rằng F là một thuộc tính trừu tượng và rằng khi các nhà lập trình nói đến “sự thân thiện”, có lẽ họ đang cố gắng truyền tải thông tin về F. Vì mục tiêu tối thượng của AI là hiện thực hóa F, tìm hiểu thêm về F sẽ là một giá trị công cụ. Khi AI càng khám phá thêm về F, hành vi của nó càng được hướng dẫn nhiều hơn bởi nội dung thực sự của F. Do đó, hy vọng rằng AI sẽ ngày càng trở nên thân thiện khi nó học được nhiều hơn và trở nên thông minh hơn.

Các lập trình viên có thể trợ giúp thêm cho quá trình này, và giảm bớt rủi ro của việc AI phạm một sai lầm nghiêm trọng nào đó trong khi sự hiểu biết về F vẫn còn chưa hoàn thiện. Điều này có thể được thực hiện bằng cách cung cấp cho AI những “lời xác nhận của lập trình viên”, nghĩa là những giả thiết về bản chất và nội dung của F được gán xác suất cao ngay từ đầu. Ví dụ, giả thiết “đánh lừa các lập trình viên là không thân thiện” có thể được gán xác suất cao ngay từ đầu. Tuy nhiên, những lời xác nhận này của lập trình viên lại không “đúng theo định nghĩa” – chúng không phải là những tiên đề luôn đúng về khái niệm thân thiện, mà thay vào đó lại là những giả thiết

ban đầu về sự thân thiện, những giả thiết mà theo đó một AI duy lý (rational AI) sẽ gán cho một xác suất cao, ít nhất cho tới khi nó vẫn còn tin vào năng lực nhận thức của lập trình viên hơn là năng lực của chính mình.

Đề xuất của Yudkowsky còn liên quan đến việc sử dụng cái mà ông gọi là “ngữ nghĩa hiệu lực nhân quả” (causal validity semantics). Ý tưởng ở đây là AI không cần làm đúng như những gì các lập trình viên đã nói, mà cần làm (thứ gì đó giống như) thứ mà họ cố gắng bảo nó làm. Trong khi các lập trình viên cố gắng giải thích cho AI hạt giống thế nào là thân thiện, họ có thể mắc sai lầm. Hơn nữa, bản thân các lập trình viên có thể cũng không hiểu đầy đủ bản chất thực sự của tính thân thiện. Do đó, người ta muốn AI có khả năng sửa lỗi trong suy nghĩ của lập trình viên và suy ra ý nghĩa đúng hoặc ý nghĩa mong muốn từ bất kỳ lời giải thích không hoàn hảo nào do lập trình viên cung cấp. Ví dụ, AI phải có khả năng hình dung được các quy luật nhân quả mà lập trình viên đã học và truyền đạt về tính thân thiện. Vì vậy, để lấy một ví dụ thông thường, AI phải hiểu được rằng một lập trình viên có thể đánh máy sai khi nhập thông tin về tính thân thiện, và phải tìm cách sửa lỗi sai này. Tổng quát hơn, AI phải tìm cách khắc phục bất cứ ảnh hưởng nào có thể bóp méo luồng thông tin về tính thân thiện khi khái niệm này được chuyển từ nguồn qua lập trình viên tới AI (trong đó “bóp méo” là một phạm trù nhận thức). Một cách lý tưởng, khi AI dần trưởng thành, nó phải vượt qua mọi thiên kiến về nhận thức và các hiểu lầm khác cơ bản hơn có thể khiến cho các lập trình viên hiểu sai về phạm trù thân thiện.

Khung 12 Hai ý tưởng hiện tại (còn non nớt)

Cách tiếp cận mà chúng ta gọi là “Hail Mary” (Chào đức mẹ Maria) dựa trên một niềm hy vọng rằng đâu đó trong vũ trụ có tồn tại (hay sẽ tồn tại) những nền văn minh đã kiểm soát thành công sự bùng nổ trí tuệ, và rằng cuối cùng họ đã có được những giá trị với độ trùng lặp đáng kể so với giá trị của chúng ta. Chúng ta có thể tìm cách xây dựng AI của mình để nó có động lực làm những gì mà các siêu trí tuệ này muốn nó thực hiện.²² Việc này dễ hơn so với việc xây dựng AI của chúng ta với động lực thực hiện những gì chúng ta mong muốn một cách trực tiếp.

Phương pháp này sẽ đạt được hiệu quả mà không cần AI của chúng ta phải thiết lập giao tiếp với bất kỳ thực thể siêu trí tuệ ngoài hành tinh nào. Thay vào đó, các hành động của AI này sẽ được hướng dẫn bởi những ước lượng của nó về những gì các siêu trí tuệ ngoài hành tinh muốn nó thực hiện. AI của chúng ta sẽ mô hình hóa các kết quả khả thể của những vụ bùng nổ trí tuệ ở nơi khác, và khi chính nó đã trở thành siêu trí tuệ thì những ước lượng đó sẽ ngày càng trở nên chính xác hơn. Kiến thức hoàn hảo là không cần thiết. Những vụ bùng nổ trí tuệ này có thể có nhiều kết quả khả thể, và AI của chúng ta khi đó sẽ cố gắng hết sức để tự điều tiết theo sở thích của nhiều loại siêu trí tuệ khác nhau có thể xuất hiện dựa vào xác suất của từng loại.

Phiên bản này của cách tiếp cận Hail Mary đòi hỏi chúng ta xây dựng một giá trị cuối cùng cho AI của mình tham chiếu đến sở thích của những siêu trí tuệ khác. Cách thức thực hiện điều này vẫn chưa thật sự rõ ràng. Tuy nhiên, các tác tử siêu trí tuệ có thể khác nhau về mặt cấu trúc tới mức chúng ta sẽ viết được một đoạn mã có chức năng như một máy dò tìm. Máy này sẽ nhìn vào mô hình thế giới

trong AI đang phát triển của chúng ta và chỉ định những phần tử đại diện tương ứng với sự hiện diện của một siêu trí tuệ. Sau đó, “cỗ máy” sẽ bằng cách nào đó rút ra những sở thích của siêu trí tuệ đang được tìm hiểu (vì nó đã được biểu diễn trong AI của chúng ta).²³ Nếu tạo ra được một máy dò như vậy, chúng ta sẽ có thể sử dụng nó để xác định những giá trị cuối cùng của AI. Thách thức ở đây là chúng ta có thể phải tạo ra máy dò trước khi biết được AI của mình sẽ phát triển khung biểu diễn nào. Vì thế, máy dò cần phải truy vấn một khung biểu diễn và kết xuất những sở thích của bất cứ siêu trí tuệ nào có thể được biểu diễn trong đó. Điều này có vẻ khó, nhưng một giải pháp thông minh nào đó có thể sẽ được tìm ra.²⁴

Nếu thiết lập cơ bản có thể hoạt động được thì những sự tinh chỉnh khác nhau ngay lập tức sẽ tự xuất hiện. Ví dụ, thay vì nhắm tới việc tuân theo (một tổ hợp có trọng số nào đó của) mọi siêu trí tuệ ngoài hành tinh, giá trị cuối cùng AI của chúng ta có thể kết hợp với một bộ lọc nhằm chọn lựa một tập hợp nhỏ các siêu trí tuệ ngoài hành tinh để làm theo (với mục tiêu lựa chọn những siêu trí tuệ có giá trị gần với chúng ta). Ví dụ, chúng ta có thể sử dụng tiêu chí đặc thù cho nguồn gốc nhân quả của một siêu trí tuệ để xác định liệu có đưa nó vào tập hợp này hay không. Một số thuộc tính nhất định về nguồn gốc của siêu trí tuệ này (mà chúng ta có thể định nghĩa được bằng các thuật ngữ cấu trúc) có thể tương quan với mức độ tương đồng giá trị được kỳ vọng của siêu trí tuệ được tạo thành. Có lẽ chúng ta mong muốn đặt thêm lòng tin vào những siêu trí tuệ có nguồn gốc nhân quả từ một bản giả lập hoàn chỉnh não bộ, hoặc từ một AI không sử dụng nhiều các thuật toán tiến hóa hoặc từ từ nổi lên như thể từ một quá trình cất cánh được kiểm soát. (Việc xem xét

những nguồn gốc nhân quả còn cho phép chúng ta tránh được việc các siêu trí tuệ được đặt trọng số quá cao có thể tạo ra các bản sao của chính mình – và chắc chắn sẽ giúp ta tránh được việc tạo động lực cho chúng làm như vậy.) Rất nhiều sự tinh chỉnh khác cũng có thể được thực hiện.

Cách tiếp cận Hail Mary đòi hỏi niềm tin rằng còn có những siêu trí tuệ khác ngoài vũ trụ có chung những giá trị với chúng ta.²⁵ Điều này khiến nó trở nên không lý tưởng. Tuy nhiên, những cản trở về kỹ thuật mà cách tiếp cận Hail Mary đang phải đối mặt tuy rất lớn nhưng chưa đáng sợ bằng những thách thức của những cách tiếp cận khác. Việc tìm tòi ra những phương pháp không lý tưởng nhưng dễ thực thi có thể có một ý nghĩa nào đó không nằm ở mục đích sử dụng, mà nằm ở phương án dự phòng trong trường hợp giải pháp lý tưởng chưa sẵn sàng được thực thi.

Một ý tưởng khác về cách giải quyết vấn đề truyền tải giá trị vừa được Paul Christiano đề xuất gần đây.²⁶ Giống như Hail Mary, đây là một phương pháp học giá trị cố gắng xác định tiêu chí giá trị bằng một “tiểu xảo” thay vì vất vả xây dựng. Ngược với Hail Mary, nó không giả định sự tồn tại của những tác tử siêu trí tuệ khác mà chúng ta có thể coi như những hình mẫu cho AI của mình. Đề xuất của Christiano khá khó để có thể giải thích ngắn gọn (nó gồm một loạt những tư duy phức tạp) nhưng ít nhất chúng ta có thể cố gắng chỉ ra một số yếu tố chính.

Giả sử chúng ta có thể có (a) một bản đặc tả chính xác về toán học của một bộ não người; và (b) một môi trường ảo được chỉ định rõ ràng về mặt toán học, trong đó có một máy tính lý tưởng với bộ nhớ và công suất CPU lớn tùy ý. Với (a) và (b), chúng ta có thể xác

định được một hàm thỏa dụng U như kết quả mà bộ não người tạo ra sau khi tương tác với môi trường này. U có thể là một đối tượng được xác định rõ ràng về mặt toán học, mặc dù chúng ta không thể xác định được nó (vì những hạn chế về tính toán). Mặc dù vậy, U có thể được dùng như tiêu chí giá trị cho một AI học giá trị có thể sử dụng những tự nghiệm khác nhau để gán xác suất cho các giả thiết về những gì mà U bao hàm.

Về trực giác, chúng ta muốn U là hàm thỏa dụng mà một người có chuẩn bị thích hợp có thể tạo ra với lợi thế sử dụng được công suất tính toán tùy ý – chẳng hạn, đủ để chạy một số lượng siêu lớn các bản sao của chính mình nhằm hỗ trợ việc phân tích để xác định một hàm tiện ích hoặc giúp đưa ra một quy trình tốt hơn để thực hiện phân tích này. (Ở đây, chúng ta đang dự báo về chủ đề “ý chí ngoại suy kết hợp” sẽ khám phá thêm trong Chương 13.)

Xem ra việc xác định môi trường lý tưởng hóa là tương đối dễ dàng: chúng ta có thể đưa ra một mô tả toán học của một máy tính trừu tượng với công suất lớn tùy ý; mặt khác, chúng ta có thể sử dụng một chương trình thực tế ảo cung cấp mô tả toán học về một phòng đơn với thiết bị đầu cuối máy tính trong đó (chính là chiếc máy tính trừu tượng). Nhưng làm thế nào để có được một mô tả chính xác về mặt toán học của một bộ não con người cụ thể? Điều này rõ ràng có thể làm được bằng giả lập hoàn chỉnh não bộ, nhưng nếu công nghệ giả lập không sẵn có thì sao?

Đây chính là điểm đổi mới quan trọng trong đề xuất của Christiano. Christiano nhận xét rằng để có được một tiêu chí giá trị được chỉ định rõ ràng về mặt toán học, chúng ta không cần một mô hình tính toán thực dụng có thể chạy được của một bộ óc nhất định

mà chỉ cần một định nghĩa toán học (có thể hàm ẩn và vô cùng phức tạp), và điều này dễ hơn rất nhiều. Với kỹ thuật hình ảnh thần kinh chức năng và các phép đo khác, có lẽ chúng ta có thể thu thập hàng gigabyte dữ liệu về hành vi đầu ra-đầu vào của một người được chọn.

Nếu chúng ta thu thập đủ lượng dữ liệu, có thể mô hình toán học đơn giản nhất tạo ra toàn bộ những dữ liệu này chính là một bản mô phỏng của người được chọn. Mặc dù không thể tính toán được mô hình này từ dữ liệu, nhưng chúng ta hoàn toàn có thể xác định được nó bằng cách tham khảo dữ liệu và sử dụng một phép đo đơn giản được xác định rõ ràng về mặt toán học (như một số biến thể của độ phức tạp Kolmogorov mà chúng ta đã gặp trong Khung 1, Chương 1).²⁷

Điều biến giả lập

Vấn đề truyền tải giá trị trên giả lập hoàn chỉnh não bộ xem ra hơi khác so với trên trí tuệ nhân tạo. Các phương pháp giả định trước sự hiểu biết và kiểm soát chi tiết về thuật toán và kiến trúc không thể áp dụng được trên các bản giả lập, nhưng phương pháp lựa chọn động lực tăng cường, vốn không áp dụng được trên trí tuệ nhân tạo ban đầu, lại có thể áp dụng trên chúng (hoặc trên bộ não sinh học đã tăng cường).²⁸

Phương pháp tăng cường này có thể được kết hợp với các kỹ thuật nhằm chỉnh sửa các mục tiêu được kế thừa của hệ thống. Chẳng hạn, người ta có thể cố gắng điều biến trạng thái động lực của một bản giả lập bằng cách cung cấp cho nó một sản phẩm kỹ thuật số có chức năng tương đương các chất kích hoạt tâm thần (hoặc chính là các hóa chất này, với một hệ thống sinh học). Thậm

chí hiện tại, việc tác động vào các giá trị và động lực cũng có thể được thực hiện bằng phương pháp dược học ở một mức độ hạn chế.²⁹ Dược điển trong tương lai có thể có các loại thuốc có tác dụng cụ thể và dễ dự đoán hơn. Môi trường kỹ thuật số của các bản giả lập cần tạo điều kiện thuận lợi cho những sự phát triển như vậy bằng cách khiến cho việc thử nghiệm có kiểm soát trở nên dễ dàng hơn và kết xuất tất cả các bộ phận khả lập địa chỉ trực tiếp của não bộ.

Giống như khi sử dụng các đối tượng thử nghiệm sinh học, nghiên cứu trên các bản giả lập sẽ gặp phải các “biến chứng” liên quan đến đạo đức, và không phải tất cả đều có thể được gạt sang một bên chỉ bằng một văn bản chấp thuận. Những vướng mắc như vậy có thể làm chậm tốc độ tiến bộ trên con đường giả lập (vì sự hiện diện của các quy định và hạn chế về đạo đức), và đặc biệt có thể cản trở các nghiên cứu về cách thức điều biến cấu trúc tạo động lực của các bản giả lập. Kết quả có thể là các bản giả lập được tăng cường đến các cấp độ siêu trí tuệ nguy hiểm tiềm tàng của khả năng nhận thức trước khi chúng ta thực hiện được đầy đủ các bước nhằm kiểm tra hay điều chỉnh các mục tiêu của chúng. Một tác động khác của các vướng mắc về đạo đức sẽ là việc “vẽ đường” cho các nhóm và quốc gia kém thận trọng hơn. Theo chiều hướng ngược lại, nếu nói lỏng các tiêu chuẩn đạo đức để thử nghiệm trí tuệ kỹ thuật số của con người, chúng ta có thể sẽ phải chịu trách nhiệm cho một lượng lớn các tác hại và việc làm sai trái, và điều này rõ ràng là không mong muốn. Khi các điều kiện khác không đổi, những cân nhắc này ủng hộ việc thực hiện một con đường thay thế nào đó

không yêu cầu sử dụng rộng rãi các đối tượng nghiên cứu kỹ thuật số là con người trong một tình huống rủi ro mang tính chiến lược.

Tuy nhiên, vấn đề này là không rõ ràng. Người ta có thể lập luận rằng việc nghiên cứu giả lập hoàn chỉnh não bộ ít có khả năng liên quan đến những vi phạm về đạo đức hơn so với việc nghiên cứu trí tuệ nhân tạo, với lý do chúng ta dễ nhận ra khi một trí óc giả lập đạt yêu cầu sở hữu vị thế đạo đức hơn so với khi một trí não hoàn toàn ngoại lai hoặc tổng hợp làm được điều này. Nếu một số loại AI nhất định hoặc các quá trình con của chúng có một vị thế đạo đức đáng chú ý mà chúng ta không nhận ra được, các hành vi vi phạm đạo đức gây ra khi đó có thể sẽ rất nghiêm trọng. Chẳng hạn, hãy xem xét việc những lập trình viên ngày nay đã tạo ra các tác tử học tăng cường một cách đầy “buông thả” và khiến chúng chịu tác động của các yếu tố kích thích phản cảm. Hằng ngày, có vô số tác tử như vậy được tạo ra, không chỉ trong các phòng thí nghiệm khoa học máy tính mà còn trong nhiều ứng dụng, bao gồm cả một số trò chơi điện tử trên máy vi tính có chứa các NPC (nhân vật không phải là người chơi) phức tạp. Có lẽ các tác tử này vẫn còn quá nguyên thủy để sở hữu được một vị thế đạo đức, nhưng làm sao chúng ta có thể thực sự tự tin rằng mọi việc xảy ra đúng như vậy? Quan trọng hơn, chúng ta có thể tự tin đến đâu về việc mình biết dừng lại đúng lúc trước khi các chương trình có được khả năng trải nghiệm được sự đau khổ liên quan đến vấn đề đạo đức?

(Trong Chương 14, chúng ta sẽ quay lại với một số câu hỏi chiến lược rộng hơn phát sinh khi so sánh mức độ đáng kỳ vọng giữa các con đường giả lập và trí tuệ nhân tạo.)

Thiết kế thể chế

Một số hệ thống thông minh có chứa các bộ phận thông minh mà bản thân chúng đều có khả năng tự chủ. Ví dụ minh họa cho điều này chính là các công ty và nhà nước của con người: trong khi đa phần do con người hợp thành, các tổ chức này có thể, vì một số mục đích, được xem như là các tác tử độc lập tự chủ. Động lực của các hệ thống đa hợp như vậy không chỉ phụ thuộc vào động lực của các tác tử con (subagent) cấu thành nên chúng, mà còn phụ thuộc vào cách thức mà các tác tử con này được tổ chức. Chẳng hạn, một nhóm được tổ chức dưới chế độ độc tài mạnh mẽ có thể hành xử như thể đang sở hữu ý chí giống hệt với ý chí của tác tử con nắm giữ vị trí độc tài, trong khi một nhóm dân chủ đôi khi lại có thể hành xử giống như thể nó có ý chí một tổng hợp hoặc trung bình của các ý chí cấu thành khác nhau. Tuy nhiên, người ta cũng có thể tưởng tượng ra các thể chế cai trị có thể khiến cho một tổ chức hành xử theo một cách thức không được tạo ra từ một hàm đơn giản (tổng hay trung bình) của ý chí của các tác tử con. (Về mặt lý thuyết, ít nhất, có thể tồn tại một nhà nước toàn trị mà mọi người đều không thích, bởi lẽ nhà nước này có cơ chế ngăn cản công dân của mình phối hợp tạo ra một cuộc nổi dậy. Người dân có thể rơi vào tình thế bất lợi nếu đơn thương độc mã chống đối mà không tham gia vào bộ máy nhà nước.)

Vì thế, bằng cách thiết kế các thể chế phù hợp cho một hệ thống đa hợp, người ta có thể cố gắng định hình động lực hiệu quả của nó. Trong Chương 9, chúng ta đã thảo luận về hòa nhập xã hội như một phương pháp kiểm soát năng lực khả thi. Nhưng ở đó, chúng ta đã tập trung vào các động lực mà một tác tử phải đối mặt do hậu quả của sự tồn tại của nó trong một thế giới xã hội gần như bị cào bằng.

Ở đây, chúng ta tập trung vào những gì xảy ra bên trong một tác tử nhất định: nghĩa là ý chí của tác tử được xác định bởi tổ chức nội bộ của chính nó như thế nào. Do đó, chúng ta sẽ xem xét một phương pháp lựa chọn động lực. Hơn nữa, vì kiểu thiết kế thể chế nội bộ này không phụ thuộc vào việc cải biến xã hội quy mô lớn, nên đây là một phương pháp có thể áp dụng cho một dự án đơn lẻ phát triển siêu trí tuệ, ngay cả khi môi trường kinh tế xã hội hoặc quốc tế rộng hơn không thuận lợi một cách lý tưởng.

Thiết kế thể chế có lẽ là phương án hợp lý nhất khi được kết hợp với phương pháp tăng cường. Nếu chúng ta có thể bắt đầu với các tác tử đã có động lực phù hợp, hoặc có động lực giống con người, thì những sự dàn xếp về mặt thể chế có thể được sử dụng như một biện pháp an toàn bổ sung để nâng cao khả năng hệ thống sẽ tiếp tục đi đúng hướng.

Giả sử, chúng ta bắt đầu với một số tác tử giống con người đã được tạo động lực phù hợp, chẳng hạn như các bản giả lập. Chúng ta muốn tăng cường năng lực nhận thức của các tác tử này, nhưng lại lo lắng rằng các biện pháp tăng cường có thể làm hỏng động lực của chúng. Một cách để đối phó với thách thức này là thiết lập một hệ thống mà trong đó, các bản giả lập riêng lẻ hoạt động như các tác tử con. Khi một biện pháp tăng cường mới được đưa vào, nó sẽ được áp dụng đầu tiên với một tập hợp tác tử con có quy mô nhỏ. Tác dụng của biện pháp tăng cường này sau đó sẽ được nghiên cứu bởi một “nhóm đánh giá” bao gồm các tác tử con chưa được áp dụng. Chỉ khi những “đồng nghiệp” này tự hài lòng rằng sự tăng cường không gây sai hỏng thì nó mới được triển khai trên nhiều tác tử con hơn. Nếu phát hiện các tác tử con đã tăng cường bị sai hỏng, chúng

sẽ không được tăng cường thêm, và sẽ được rút khỏi các chức năng ra quyết định quan trọng (ít nhất là cho đến khi toàn bộ hệ thống tiến đến một điểm mà ở đó các tác tử con bị sai hỏng có thể được tái hòa nhập một cách an toàn).³⁰ Tuy các tác tử con sai hỏng có thể đã đạt được một số lợi thế từ việc tăng cường, nhưng cấu trúc thể chế mà chúng thuộc về, cũng như việc chúng tạo thành một nhóm tác tử con thiếu số, sẽ hoàn toàn có thể ngăn chúng chiếm quyền kiểm soát hoặc phát tán lỗi sai hỏng ra toàn bộ hệ thống. Do đó, trí tuệ tập thể và khả năng của hệ thống có thể được tăng cường dần dần theo một chuỗi các bước nhỏ, trong đó mức độ hoàn hảo của từng bước được xác minh bởi các tác tử con chỉ có khả năng kém hơn một chút so với các tác tử con mới được đưa vào trong bước đó.

Mức độ an toàn có thể đạt được của kiểu thiết kế tổ chức này bị giới hạn bởi độ chính xác của các thử nghiệm được sử dụng để đánh giá các tác tử con được tăng cường. Vấn đề xây dựng thử nghiệm là rất phức tạp, vì các chuỗi thử nghiệm (test batteries) chỉ có thể được xác nhận trong phạm vi khả năng nhận thức đã có từ trước nhưng lại phải được áp dụng cho các tác tử nằm ngoài phạm vi này. Người ta có thể giảm thiểu vấn đề này bằng cách thực hiện các bước tăng cường nhỏ và để thử nghiệm chạy trong một thời gian dài. Tuy nhiên, sự thận trọng như vậy sẽ làm tăng chi phí và chậm tiến độ. (Điều này đồng nghĩa với việc một dự án áp dụng các biện pháp an toàn này sẽ tự đặt mình vào tình thế bất lợi khi xuất hiện một động lực chạy đua nào đó.) Ngay cả khi thực hiện các bước nhỏ và thử nghiệm kéo dài, vẫn sẽ tồn tại một rủi ro về việc các lỗi tinh vi sẽ xuất hiện, dẫn đến việc hệ thống mục tiêu bị sai hỏng dần dần. Sự sai hỏng này chỉ biểu hiện ra ngoài khi hệ thống đạt đến siêu trí tuệ. Rủi

ro này có thể được giảm bớt bằng cách kéo dài tình trạng thử nghiệm thành nhiều bước nâng cao nhận thức, sao cho ngay cả các tác tử vốn ban đầu được xác định là hoàn hảo và đã nhận được một số sự tăng cường tiếp theo sẽ được xem xét bởi một nhóm tác tử gốc chưa được sửa đổi và chưa sai hỏng, và sẽ vẫn có thể bị đảo ngược quá trình.

Lý tưởng nhất, việc kiểm tra nên được thực hiện liên tục chứ không bị giới hạn ở một giai đoạn đánh giá ban đầu. Các tác tử con ở cấp độ con người (là con người) sở hữu khả năng lập chiến lược, và do đó có thể chọn cách che giấu một số mục tiêu nhất định trong khi hành vi của mình đang bị theo dõi. Để tiến hành thử nghiệm liên tục, một hệ thống phân cấp có thể được tạo ra, trong đó các tác tử con có một cấp độ năng lực nhất định được giao nhiệm vụ giám sát các tác tử con ở một cấp độ năng lực nào đó cao hơn. Ở dưới cùng của hệ thống phân cấp năng lực (nhưng ở trên cùng của hệ thống phân cấp quyền lực) sẽ là người đứng đầu nhưng tương đối ngốc nghếch và chậm chạp. Người đứng đầu này sẽ giống như một vị vua mất trí đứng đầu một triều đình thiếu năng lực, và “triều đình” này lại đang quản lý một bộ máy chính quyền tầm thường nhưng cai trị những người dân xuất chúng. Ngoài ra, người này có thể được ví như bộ não bò sát đã trở nên cổ xưa qua quá trình tiến hóa, với nhiệm vụ đảm bảo rằng tân vỏ não mới hơn và thông minh hơn được sử dụng để phục vụ các mục tiêu như ăn uống và tình dục.

Người ta sẽ đặt câu hỏi rằng liệu một chế độ nhân tài “ngược” như vậy có thể duy trì ổn định. Nếu phần lớn các tác tử con siêu trí tuệ bị phát hiện là sai hỏng, và những người chịu trách nhiệm mọi việc đã cố gắng đảo ngược rất nhiều biện pháp tăng cường đã thực

hiện, thì liệu cấu trúc thể chế này có đủ mạnh để chống lại sự phản đối mà một sự vận động như vậy sẽ gây ra? Ở bối cảnh xã hội loài người trong lịch sử hoặc đương đại, có lẽ chúng ta phải trả lời rằng điều này là không thể. (Những người có khả năng nhất không phải lúc nào cũng vươn lên nắm giữ quyền lực, nhưng chúng ta cũng không bị tinh tinh cai trị theo nghĩa đen.) Tuy nhiên, vấn đề sẽ không rõ ràng như vậy trong bối cảnh xã hội được vận hành như một phần mềm.³¹

Hãy tưởng tượng một chế độ phân cấp kỹ thuật số mà trong đó, cứ hai người vô sản lại được giám sát bởi một cảnh sát được đào tạo cấp cao, hai cảnh sát được giám sát bởi một cảnh sát đặc biệt, hai cảnh sát đặc biệt lại được giám sát bởi một cảnh sát đặc biệt khác, và cứ thế; sao cho mỗi cấp độ giám sát đều có số lượng bằng một nửa cấp độ được giám sát tương ứng. Hơn nữa, hãy tưởng tượng rằng xã hội này được kiến trúc theo cách mà người giám sát có lợi thế lớn so với cấp dưới mình. Ví dụ, người giám sát có thể theo dõi và ghi lại tất cả những gì cấp dưới nói và làm, trong khi bản thân người giám sát lại được che giấu sau lớp kính một chiều, sao cho cấp dưới của anh ta chỉ có thể nghe được những mệnh lệnh và hướng dẫn mà người cấp trên lựa chọn để truyền đạt. Hãy tưởng tượng các giám sát viên có một bảng điều khiển mà họ có thể ấn bất cứ lúc nào để khiến cấp dưới bị phạt, khen thưởng, tạm dừng, xóa hoặc trở về một trạng thái trước đó. Những năng lực này sẽ thúc đẩy đáng kể sức mạnh của một giám sát viên, vượt xa những gì khả thi trong lịch sử ngay cả trong những tổ chức toàn trị nhất.

Ngay cả điều này cũng không làm cạn kiệt các công cụ kiểm soát khả thể. Các giám sát viên cũng sở hữu các bản mô phỏng thực tế

ảo thực sự có thể dùng để thăm dò phản ứng của cấp dưới đối với các tình huống giả định (chẳng hạn như các cơ hội mô phỏng để tham gia một âm mưu hoặc thả lỏng khi không có ai giám sát). Ngoài ra, các giám sát viên có thể có khả năng nghe lén các đoạn độc thoại nội tâm trong tâm trí cấp dưới và trực tiếp thao túng trạng thái cảm xúc của họ.

Kết quả cuối cùng là giám sát viên sẽ có một loạt các lợi thế lớn.³²

Mỗi giám sát viên sẽ là cấp dưới của một giám sát viên cấp cao khác, người sẽ giám sát mọi hành động của anh ta và đảm bảo rằng anh ta không bị chậm trễ trong việc thực hiện các nhiệm vụ quản lý của mình. Với kiểu sắp xếp này, cấu trúc có thể ổn định ngay cả khi mỗi cấp độ giám sát yếu hơn về mặt trí tuệ so với cấp độ được giám sát tương ứng.³³

Và toàn bộ cấu trúc, có thể bao gồm nhiều tác tử siêu trí tuệ, có thể được kiểm soát bởi một số ít người có quyền truy cập root (root access).³⁴

Chi phí của một cấu trúc kiểm soát nội bộ như vậy sẽ rất cao, nhưng không hẳn là không thể chi trả. Chẳng hạn, trong mô hình mà mỗi cấp độ giám sát có số cá thể bằng một nửa cấp độ bên dưới, chi phí tính toán bổ sung được giới hạn ở mức 100% chi phí phải trả cho phần “vô sản” của hệ thống, và sẽ thấp hơn nếu các cấp độ “ông chủ” ngốc nghếch ở trên đòi hỏi số lượng tính toán ít hơn với mỗi tác tử con. Việc giám sát có thể được thắt chặt khi tăng tỷ lệ này, hoặc giảm chi phí bằng cách giảm tỷ lệ. Chi phí tính toán bổ sung sẽ là một yếu tố quan trọng trong kịch bản thị trường cạnh tranh (xem Chương 11), nhưng xem ra sẽ chấp nhận được nếu dự án không phải lập tức đối

mặt với sự cạnh tranh mạnh mẽ. Dự án cũng sẽ phát sinh một loại chi phí về khoảng thời gian phát triển cần thiết để khởi tạo và thử nghiệm các chức năng giám sát này. Một dự án có đủ tài nguyên có thể giảm khoản chi phí thời gian này bằng cách vận hành quá trình phát triển cấu trúc kiểm soát song song với việc phát triển trí tuệ máy; nhưng những nhiệm vụ bổ sung có thể sẽ quá đắt đỏ với các dự án quy mô nhỏ hơn và các dự án đang cuốn vào một cuộc đua công nghệ sít sao.

Một loại chi phí khác cũng đáng được xem xét, đó là rủi ro của việc những hành vi tội ác tâm trí được thực hiện trong loại cấu trúc này.³⁵ Theo những gì được mô tả, thể chế này nghe giống như một trại tập trung đáng sợ. Tuy nhiên, có nhiều cách để ít nhất có thể giảm nhẹ các vấn đề đạo đức khi điều hành loại thể chế này, ngay cả khi các tác tử con trong tổ chức là các bản giả lập có đầy đủ vị thế đạo đức như con người. Ở một mức tối thiểu, hệ thống có thể dựa vào các bản giả lập tình nguyện. Mỗi tác tử có thể tùy ý lựa chọn dừng tham gia ở bất kỳ thời điểm nào.³⁶ Các bản giả lập dừng tham gia có thể được lưu vào bộ nhớ, với một sự đảm bảo về việc khởi động lại chúng trong những điều kiện lý tưởng hơn nhiều, khi giai đoạn nguy hiểm của sự bùng nổ trí tuệ đã qua đi. Trong khi đó, tác tử con chọn tham gia có thể được ở trong những môi trường ảo rất thoải mái và cho phép có nhiều thời gian để nghỉ ngơi và giải trí. Những biện pháp này sẽ tạo ra một khoản chi phí có thể chấp nhận được đối với một dự án có nguồn lực đầy đủ trong điều kiện không cạnh tranh. Trong tình huống cạnh tranh cao, chi phí này có thể sẽ không chi trả được, trừ khi doanh nghiệp đảm bảo được rằng các đối thủ cạnh tranh sẽ phải chịu chi phí tương tự.

Trong ví dụ này, chúng tôi tưởng tượng rằng các tác tử con là các bản giả lập. Người ta có thể tự hỏi, liệu phương pháp thiết kế tổ chức có đòi hỏi tác tử con phải là nhân hình? Hoặc liệu nó có thể áp dụng tương đương cho các hệ thống bao gồm các tác tử con nhân tạo?

Suy nghĩ đầu tiên của ai đó ở đây có thể sẽ là hoài nghi. Ta nhận ra rằng tuy có nhiều kinh nghiệm với các tác tử giống con người, nhưng chúng ta vẫn không thể dự đoán chính xác sự bùng phát hoặc kết quả của các cuộc cách mạng; khoa học xã hội chỉ có thể mô tả được một số xu hướng mang tính thống kê mà thôi.³⁷ Chúng ta không thể dự đoán một cách đáng tin cậy sự ổn định của các cấu trúc xã hội với con người bình thường (là các “tác tử” mà chúng ta có nhiều dữ liệu). Điều này sẽ dẫn tới suy luận rằng chúng ta có rất ít hy vọng chế tạo một cách chính xác các cấu trúc xã hội ổn định cho các tác tử giống con người đã được tăng cường nhận thức (mà chúng ta không hề có dữ liệu), và thậm chí còn có ít hy vọng hơn về việc làm được điều này cho các tác tử nhân tạo tiên tiến (vốn không hề giống với các tác tử mà chúng ta có dữ liệu).

Tuy nhiên, vấn đề không hẳn đã được quyết định. Con người và các cá thể giống con người là rất phức tạp, nhưng các tác tử nhân tạo có thể lại sở hữu các kiểu kiến trúc tương đối đơn giản. Các tác tử nhân tạo cũng có thể có các động lực đặc trưng đơn giản và tường minh. Hơn nữa, các tác tử kỹ thuật số nói chung (dù là giả lập hay trí tuệ nhân tạo) đều có thể sao chép, và khả năng này có thể cách mạng hóa quá trình quản lý, giống như các bộ phận có thể hoán đổi đã cách mạng hóa qua trình sản xuất vậy. Những khác biệt này, cùng với cơ hội làm việc với các tác tử ban đầu không sở hữu năng

lực và cơ hội tạo ra các cấu trúc thể chế sử dụng các biện pháp kiểm soát khác nhau đã nêu phía trên, có thể kết hợp nhằm có được những kết quả cụ thể về thể chế (chẳng hạn như một hệ thống không chống đối) đáng tin cậy hơn so với khi làm việc cùng con người trong các điều kiện lịch sử.

Nhưng một lần nữa, các tác tử nhân tạo có thể sẽ thiếu đi nhiều thuộc tính giúp chúng ta dự đoán hành vi như ở các tác tử giống con người. Các tác tử nhân tạo không có bất kỳ cảm xúc xã hội nào gắn với hành vi, cảm xúc của con người, như sợ hãi, tự hào và hối hận, và cũng không phát triển sự gắn kết với bạn bè và gia đình. Chúng cũng không thể hiện thứ ngôn ngữ cơ thể vô thức đã khiến con người khó lòng che giấu được chủ ý. Những sự thiếu hụt này có thể làm mất tính ổn định của các thể chế của tác tử nhân tạo. Hơn nữa, các tác tử nhân tạo có thể có khả năng tạo ra bước nhảy vọt lớn về hiệu năng nhận thức do kết quả của những thay đổi dường như rất nhỏ về thuật toán hoặc kiến trúc của chúng. Các tác tử nhân tạo tối ưu hóa một cách không khoan nhượng có thể sẵn sàng chấp nhận những canh bạc mạo hiểm mà con người không thể.³⁸ Bên cạnh đó, các tác tử siêu trí tuệ có thể cho thấy khả năng phối hợp đáng ngạc nhiên mà hầu như không hề giao tiếp (ví dụ: bằng cách mô hình hóa từ bên trong các phản ứng mang tính giả thuyết của nhau đối với các tình huống khác nhau). Những điều vừa nêu, cùng những sự khác biệt khác, có thể làm gia tăng khả năng của một sự thất bại bất ngờ về thể chế ngay cả khi có sự hiện diện của những biện pháp kiểm soát xã hội bền vững.

Do đó, cách tiếp cận thiết kế thể chế hứa hẹn như thế nào, và liệu nó có nhiều cơ hội thành công trên các tác tử giống con người hơn là

trên các tác tử nhân tạo hay không là không rõ ràng. Ta có thể nghĩ rằng việc tạo ra một thể chế có hoạt động kiểm tra và cân đối (checks and balances) phù hợp chỉ có thể làm tăng mức độ an toàn (hoặc không làm giảm mức độ an toàn trong bất kỳ trường hợp nào) sao cho từ góc độ giảm thiểu rủi ro, mọi thứ sẽ luôn ở trong điều kiện tốt nhất nếu phương pháp này được sử dụng. Nhưng ngay cả điều này cũng không thể khẳng định chắc chắn. Cách tiếp cận này bổ sung thêm nhiều thành phần và làm tăng độ phức tạp, và vì thế cũng có thể khiến mọi thứ bị xáo trộn theo những đường hướng mới vốn không tồn tại nếu tác tử không được cấu thành từ các tác tử con có trí tuệ. Nhưng dù sao đi nữa, phương pháp thiết kế thể chế cũng đáng để chúng ta nghiên cứu sâu hơn.³⁹

Tóm tắt

Xác định hệ thống mục tiêu (goal system engineering) chưa phải là một ngành chính thức. Hiện tại, chúng ta vẫn chưa biết được cách thức truyền tải giá trị của con người sang một máy tính kỹ thuật số, ngay cả khi đã có trí tuệ máy cấp độ con người. Sau khi nghiên cứu một số cách tiếp cận, chúng tôi thấy rằng dường như vẫn có ít nhiều ngõ cụt, nhưng một số khác lại đầy hứa hẹn và cần được khám phá thêm. Bảng 12 sẽ đưa ra những nội dung tóm lược.

Khung tạo động lực	Vẫn còn quá sớm để nói về mức độ khó khăn của việc khuyến khích một hệ thống phát triển các phép biểu diễn nội bộ cấp cao mà con người dễ nhìn thấy (trong khi vẫn giữ được năng lực của hệ thống dưới mức nguy hiểm), sau đó sử dụng các phép biểu diễn này để thiết kế một hệ thống mục tiêu mới. Cách tiếp cận này có thể tương đối hứa hẹn. (Tuy nhiên, cũng giống như bất kỳ phương pháp chưa được thử nghiệm nào sẽ trì hoãn quá trình nghiên cứu kỹ thuật an toàn cho đến khi AI cấp độ con người được phát triển, ta cần thận trọng tránh việc phương pháp này trở thành một lời biện hộ cho thái độ “yếu đuối” trước vấn đề kiểm soát trong giai đoạn chuyển tiếp.)
Học giá trị	Một cách tiếp cận đầy tiềm năng, nhưng cần nghiên cứu thêm để xác định mức độ khó khăn khi chỉ định chính thức một tham chiếu thành công trong việc chỉ ra những thông tin bên ngoài có liên quan về giá trị của con người (và khi chỉ định một tiêu chí đúng sai cho hàm thỏa dụng về một tham chiếu như vậy). Trong phạm trù học giá trị, chúng ta cũng nên khám phá về phương pháp Hail Mary hoặc đi theo phương pháp xây dựng của Paul Christiano (hay các lối tắt tương tự khác).
Điều biến giả lập	Nếu trí tuệ máy đạt được thông qua con đường giả lập, rất có thể chúng ta sẽ điều chỉnh được các động lực thông qua các loại “thuốc” kỹ thuật số hoặc các biện pháp khác. Liệu điều này có cho phép các giá trị được truyền tải với độ chính xác đủ để đảm bảo an toàn ngay cả khi việc mô phỏng được tăng cường đến siêu trí tuệ hay không vẫn là một câu hỏi mở. (Những hạn chế về đạo đức cũng có thể làm phức tạp thêm quá trình phát triển theo hướng này.)
Thiết kế thể chế	Nhiều phương pháp kiểm soát xã hội mạnh mẽ có thể được áp dụng trong một thể chế tạo thành bởi các bản giả lập. Về nguyên tắc, các phương pháp kiểm soát xã hội cũng có thể được áp dụng trong một thể chế của các tác tử con trí tuệ nhân tạo. Các bản giả lập một số thuộc tính giúp chúng dễ bị kiểm soát hơn thông qua các phương pháp như vậy, nhưng cũng có một số thuộc tính có thể khiến chúng khó bị kiểm soát hơn AI. Thiết kế thể chế là phương pháp đáng để nghiên cứu sâu hơn như một kỹ thuật truyền tải giá trị đầy tiềm năng.

Khi đã biết cách giải quyết vấn đề truyền tải giá trị, chúng ta sẽ phải đối mặt với một vấn đề khác: quyết định giá trị cần truyền tải. Nói cách khác, đó là câu hỏi về thứ mà chúng ta muốn một siêu trí tuệ phải muốn. Đây chính là vấn đề có tính triết học cao hơn sẽ được bàn tới trong phần tiếp theo.

Chương 13

Lựa chọn tiêu chí chọn lựa

Gia sử chúng ta có thể cài đặt bất kỳ giá trị cuối cùng tùy ý nào vào một AI hạt giống, thì quyết định về việc giá trị nào được cài đặt sau đó sẽ có thể gây ra những hậu quả sâu rộng hơn cả. Một số lựa chọn tham số cơ bản khác, liên quan đến các tiên đề của lý thuyết quyết định và nhận thức luận của AI, cũng có thể gây ra hậu quả tương tự. Nhưng làm thế nào ta tin tưởng được rằng những bộ óc gốc ngoặc và hạn hẹp của mình sẽ đưa ra những quyết định thiết kế sáng suốt? Làm sao chúng ta có thể lựa chọn mà không cần khóa chặt vĩnh viễn những định kiến của thế hệ hiện tại? Trong chương này, chúng ta sẽ khám phá những cách thức sao cho quy chuẩn gián tiếp có thể cho phép truyền tải phần lớn công việc nhận thức liên quan đến việc đưa ra các quyết định này sang siêu trí tuệ, trong khi vẫn giới hạn được kết quả trong phạm vi các giá trị sâu hơn của con người.

Nhu cầu quy chuẩn

Làm sao để một siêu trí tuệ làm những gì mà chúng ta mong muốn? Chúng ta cần siêu trí tuệ mong muốn điều gì? Từ đầu đến giờ, chúng ta đã tập trung vào câu hỏi thứ nhất, và bây giờ sẽ giải quyết câu hỏi thứ hai.

Giả sử chúng ta đã giải quyết vấn đề kiểm soát để có thể truyền tải bất kỳ giá trị đã chọn nào lên hệ thống động lực của một siêu trí tuệ, và khiến nó theo đuổi giá trị đó như mục tiêu tối thượng. Vậy khi đó, chúng ta nên cài đặt giá trị nào? Ở đây, lựa chọn không phải là một vấn đề đơn giản. Nếu siêu trí tuệ có được một ưu thế chiến lược quyết định, giá trị đó sẽ quyết định thiên hướng của tài nguyên vũ trụ.

Rõ ràng, chúng ta không được phạm sai lầm trong quá trình chọn lọc giá trị của mình. Nhưng làm thế nào chúng ta thực sự có thể hy vọng rằng mình không gây ra lỗi sai trong một vấn đề như vậy? Chúng ta có thể sai về đạo đức; sai về những gì tốt cho mình; thậm chí sai về những gì chúng ta thực sự muốn. Xem ra, việc chỉ định một mục tiêu tối thượng đòi hỏi chúng ta phải cẩn thận xem xét một loạt các vấn đề triết lý gai góc. Nếu thử áp dụng một cách tiếp cận trực tiếp, chúng ta sẽ dễ tạo ra một mớ bòng bong. Nguy cơ lựa chọn sai lầm sẽ trở nên đặc biệt cao khi bối cảnh quyết định là không quen thuộc, và việc chọn mục tiêu tối thượng cho một siêu trí tuệ máy định hình toàn bộ tương lai của loài người sẽ là một bối cảnh quyết định vô cùng lạ lẫm (nếu tồn tại).

Tỷ lệ ủng hộ ảm đảm với cách thức “tấn công trực diện” này được phản ánh trong sự bất đồng tràn lan về các vấn đề có liên quan trong lý thuyết giá trị: Không lý thuyết đạo đức nào nhận được sự ủng hộ đa số từ các triết gia, vì thế hầu hết các triết gia sai.¹ Điều này cũng được phản ánh qua những thay đổi rõ rệt trong sự phân bố niềm tin đạo đức theo thời gian, và chúng ta luôn muốn nghĩ rằng nhiều thay đổi trong đó là tiến bộ. Ở châu Âu thời trung cổ, việc chứng kiến một tù nhân chính trị bị tra tấn đến chết được coi là một trò tiêu khiển, và việc đốt cháy những chú mèo vẫn còn phổ biến ở

Paris vào thế kỷ 16.² Chỉ mới 150 năm trước đây, nô lệ vẫn còn là một vấn đề phổ biến ở Nam Mỹ và được pháp luật cũng như truyền thống đạo đức ủng hộ hoàn toàn. Khi nhìn lại, chúng ta thấy được những thiếu sót rõ ràng không chỉ trong hành vi mà cả niềm tin đạo đức của mọi thời đại trước. Mặc dù có thể đã lược lặt được một số kiến thức về đạo đức, chúng ta khó có thể tuyên bố rằng mình đang chìm trong ánh sáng của sự giác ngộ đạo đức hoàn hảo. Rất có thể, chúng ta vẫn đang lao động dưới sự chi phối của một hoặc nhiều quan niệm sai lầm nghiêm trọng về đạo đức. Trong những trường hợp như vậy, việc lựa chọn một giá trị cuối cùng dựa trên niềm tin hiện tại của chúng ta, bằng cách vĩnh viễn “khóa chặt” nó và ngăn chặn mọi khả năng xảy ra sự tiến bộ về đạo đức, sẽ có nguy cơ gây ra một thảm họa đạo đức sống còn.

Ngay cả khi chúng ta có thể tin tưởng một cách hợp lý rằng mình đã xác định được lý thuyết đạo đức chính xác (việc mà chúng ta không thể), chúng ta vẫn có nguy cơ mắc sai lầm trong việc phát triển các chi tiết quan trọng của lý thuyết này. Các lý thuyết đạo đức đơn giản xem ra cũng có thể có rất nhiều điều phức tạp còn ẩn giấu.³ Chẳng hạn, hãy xem xét lý thuyết hệ quả luận theo chủ nghĩa khoái lạc (một lý thuyết rất đơn giản). Lý thuyết này tuyên bố, một cách đại khái, rằng tất cả các khoái lạc đều tốt đẹp, còn toàn bộ nỗi đau đều xấu xa. Ngay cả khi chúng ta đặt toàn bộ các “con chip đạo đức” của mình vào lý thuyết này, và lý thuyết này đã đúng, rất nhiều câu hỏi sẽ vẫn còn bỏ ngỏ. Những “niềm khoái lạc cao hơn” có nên được ưu tiên so với những “niềm khoái lạc thấp hơn”, như John Stuart Mill đã tranh luận? Chúng ta nên tính đến cường độ và thời gian của một niềm khoái lạc như thế nào? Đau đớn và khoái lạc có thể loại trừ lẫn

nhau hay không? Những loại trạng thái nào của não bộ có liên quan đến những khoái lạc về đạo đức? Liệu sở hữu hai bản sao chính xác của cùng một trạng thái não bộ có tương ứng với việc có khoái lạc gấp đôi?⁵ Liệu có tồn tại những khoái lạc trong tiềm thức? Chúng ta nên xử trí ra sao trước những cơ hội rất nhỏ của những khoái lạc vô cùng lớn?⁶ Chúng ta nên tiến hành tổ hợp ra sao trên một quần thể vô hạn?⁷

Trả lời sai bất kỳ câu hỏi nào trên đây đều có thể là thảm họa. Nếu khi chọn một giá trị cuối cùng cho siêu trí tuệ, chúng ta phải đặt cược không chỉ vào một lý thuyết đạo đức chung, mà cả một dải liên hợp dài những tuyên bố cụ thể về cách lý thuyết đó được diễn giải và tích hợp vào một quá trình ra quyết định hiệu quả, cơ hội may mắn của chúng ta sẽ giảm xuống một mức gần như vô vọng. Kẻ ngốc có thể háo hức chấp nhận thử thách giải quyết tất cả các vấn đề quan trọng trong triết lý đạo đức, nhằm đưa các câu trả lời yêu thích của họ vào AI hạt giống. Ngược lại, những người thông thái sẽ gắng sức tìm kiếm một phương pháp tiếp cận thay thế, một cách thức “rào giậu”.

Điều này dẫn chúng ta đến vấn đề về quy chuẩn gián tiếp. Lý do rõ ràng của việc xây dựng một siêu trí tuệ là để có thể truyền tải cho nó lý do công cụ cần có, nhằm tìm ra những phương án hiệu quả trong việc hiện thực hóa một giá trị cho trước. Quy chuẩn gián tiếp sẽ cho phép chúng ta truyền tải cho siêu trí tuệ lý do cần thiết để chọn ra giá trị sẽ được hiện thực hóa.

Quy chuẩn gián tiếp là một cách đáp lại thách thức được đưa ra bởi việc chúng ta có thể không biết mình thực sự muốn gì, những gì là lợi ích của chúng ta, hoặc những gì là đúng đắn hoặc lý tưởng về

đạo đức. Thay vì đưa ra dự đoán dựa trên hiểu biết hiện tại của mình (có lẽ là thiếu sót sâu sắc), chúng ta sẽ ủy thác một phần công việc nhận thức cần thiết cho việc chọn lọc giá trị cho siêu trí tuệ. Siêu trí tuệ giỏi hơn chúng ta trong hoạt động nhận thức, vì thế, nó có thể nhìn thấy những lỗi sai và sự nhầm lẫn che mờ suy nghĩ của chúng ta. Ý tưởng này có thể được khái quát và nhấn mạnh như một nguyên tắc tự nghiệm:

Nguyên tắc tôn trọng nhận thức (epistemic deference)

Một siêu trí tuệ trong tương lai chiếm một điểm thuận lợi vượt trội về mặt nhận thức: những gì nó tin tưởng (có lẽ, về hầu hết các chủ đề) sẽ có nhiều khả năng đúng hơn so với niềm tin của chúng ta. Do đó, chúng ta nên tuân theo quan điểm của siêu trí tuệ bất cứ khi nào có thể.⁸

Quy chuẩn gián tiếp áp dụng nguyên tắc này cho vấn đề chọn lọc giá trị. Khi không tin tưởng rằng bản thân mình có thể xác định một tiêu chuẩn quy chuẩn cụ thể, chúng ta sẽ chỉ định một số điều kiện trừu tượng hơn mà bất kỳ tiêu chuẩn quy chuẩn nào cũng phải thỏa mãn, với hy vọng rằng một siêu trí tuệ có thể tìm thấy một tiêu chuẩn cụ thể thỏa mãn các điều kiện trừu tượng này. Chúng ta có thể cung cấp cho AI hạt giống mục tiêu tối thượng là tiếp tục hành động theo sự ước tính khả quan nhất của nó về những gì mà tiêu chuẩn được định nghĩa ngầm này sẽ đòi hỏi nó thực hiện.

Một số ví dụ sẽ làm rõ ý tưởng này. Trước tiên, chúng ta sẽ xem xét vấn đề “ý chí ngoại suy kết hợp”, một đề xuất về quy chuẩn gián tiếp của Eliezer Yudkowsky. Sau đó, một số biến thể và phương án thay thế sẽ được giới thiệu, qua đó, chúng ta sẽ biết được một số lựa chọn sẵn có.

Ý chí ngoại suy kết hợp

Yudkowsky đã đề xuất rằng một AI hạt giống sẽ được đưa ra mục tiêu tối thượng là thực hiện ý chí ngoại suy kết hợp (coherent extrapolated volition, CEV) của con người. Ông định nghĩa CEV như sau:

Ý chí ngoại suy kết hợp là điều mà chúng ta mong muốn nếu mình biết nhiều hơn, suy nghĩ nhanh hơn, trở thành những hình mẫu mà mình mong muốn, hay đã cùng nhau trưởng thành hơn. Trong điều mong muốn đó, phép ngoại suy là hội tụ chứ không phân tách, những mong muốn của chúng ta là kết hợp chứ không đối lập. Điều mong muốn đó sẽ được ngoại suy như cách ta muốn nó được ngoại suy, và được giải thích như cách ta muốn nó được giải thích.⁹

Khi viết định nghĩa trên, Yudkowsky không định trình bày kế hoạch chi tiết về cách thực hiện “toa thuốc” đầy thi vị này. Mục đích của ông là đưa ra một bản phác thảo sơ bộ về cách định nghĩa CEV, cùng với một số lập luận về lý do tại sao ta cần một cách tiếp cận như vậy.

Nhiều ý tưởng đằng sau đề xuất về CEV có những điểm tương đồng và tiền đề trong các tài liệu triết học. Chẳng hạn, trong đạo đức học, các lý thuyết về người quan sát lý tưởng tìm cách phân tích các khái niệm quy chuẩn như “tốt”, hay “đúng” theo những nhận xét mà một người quan sát lý tưởng giả định sẽ đưa ra (trong đó, người quan sát lý tưởng giả định được định nghĩa là một người không hề biết đến những thông tin phi đạo đức, có cái nhìn logic rõ ràng, là người vô tư theo những cách phù hợp, hoàn toàn không có thiên kiến,...).¹⁰ Tuy nhiên, cách tiếp cận CEV không phải (hoặc không cần phải được hiểu) là một lý thuyết đạo đức. Sẽ không đảm bảo khi nói

rằng có bất kỳ mối liên hệ cần thiết nào giữa giá trị và những điểm ưu tiên của ý chí ngoại suy kết hợp. CEV có thể đơn giản được coi là một phương pháp hữu ích để ước lượng bất cứ thứ gì sở hữu một giá trị cuối cùng. Hoặc, khái niệm này cũng có thể được xem xét bên ngoài bất kỳ mối liên hệ nào với đạo đức học. Là nguyên mẫu chính của phương pháp quy chuẩn gián tiếp, khái niệm này rất đáng để chúng ta xem xét chi tiết hơn.

Một vài chú giải

Chúng ta cần phải giải thích một số từ ngữ trong phần trích dẫn trên. “Suy nghĩ nhanh hơn” trong thuật ngữ của Yudkowsky là nếu chúng ta thông minh hơn và suy nghĩ nhiều hơn. “Cùng nhau trưởng thành hơn” xem ra nghĩa là nếu chúng ta đã học hỏi, nâng cao nhận thức và tự cải thiện bản thân trong các điều kiện tương tác xã hội phù hợp với nhau.

“Phép ngoại suy là hội tụ chứ không phân tách” có thể hiểu như sau: AI nên hành động dựa trên một số tính năng của kết quả tự ngoại suy chỉ trong phạm vi mà tính năng đó có thể được nó dự đoán với độ tin cậy khá cao. Trong phạm vi mà AI không thể dự đoán những gì chúng ta mong muốn nếu chúng ta được lý tưởng hóa theo cách đã nêu, AI không nên hành động theo phỏng đoán tự nhiên, mà thay vào đó nên tự kiểm chế. Tuy nhiên, mặc dù nhiều chi tiết về mong muốn đã được lý tưởng hóa của chúng ta có thể không được xác định hoặc không thể đoán trước, vẫn có thể có một số nét chính mà AI có thể nắm bắt, và sau đó, ít nhất nó có thể hành động để đảm bảo rằng các sự kiện trong tương lai diễn ra trong khuôn khổ các nét chính đó. Ví dụ, nếu có thể ước tính một cách đáng tin cậy rằng ý chí ngoại suy của chúng ta mong muốn tất cả chúng ta không liên tục

chìm trong đau khổ, hoặc vũ trụ không bị xếp chồng lên nhau bằng kẹp giấy, thì AI nên hành động để ngăn chặn những kết quả đó.¹¹

“Những mong muốn của chúng ta là kết hợp chứ không đối lập” có thể được hiểu như sau: AI nên hành động khi có sự đồng thuận giữa hầu hết các cá nhân. Một tập hợp nhỏ những mong muốn mạnh mẽ và rõ ràng đôi khi có thể vượt xa những mong muốn yếu ớt và hỗn độn của đa số. Ngoài ra, Yudkowsky nghĩ rằng sự đồng thuận để AI ngăn chặn một số kết quả được chỉ định cụ thể cần giảm xuống, còn sự đồng thuận để AI hành động nhằm đưa tương lai vào một khái niệm hẹp cụ thể về cái “tốt” lại phải tăng thêm. Ông đã viết: “Động lực ban đầu cho CEV nên là thận trọng khi nói ‘đúng’, và lắng nghe cẩn thận những điều ‘không đúng’.”¹²

“Điều mong muốn đó sẽ được ngoại suy như cách ta muốn nó được ngoại suy, và được giải thích như cách ta muốn nó được giải thích”: Ý tưởng đằng sau nội dung cuối cùng này dường như là việc bản thân các quy tắc ngoại suy nên nhạy cảm với ý chí ngoại suy. Một cá nhân có thể có một “mong muốn cấp hai” (một mong muốn liên quan đến điều được mong muốn) rằng một ham muốn cấp một nào đó của mình sẽ không còn quan trọng khi ý chí được ngoại suy. Ví dụ, một người nghiện rượu có ham muốn thứ nhất là được ăn nhậu say sưa cũng có thể có ham muốn thứ hai rằng mình không có ham muốn thứ nhất đó. Tương tự, chúng ta có thể có mong muốn về việc các cấu phần khác của quá trình ngoại suy sẽ diễn ra như thế nào, và những mong muốn này cũng cần được quá trình ngoại suy tính đến.

Nhiều người có thể sẽ phản bác rằng ngay cả khi khái niệm về ý chí ngoại suy kết hợp của con người có thể được định nghĩa chính

xác, thì dù gì đi nữa, việc tìm hiểu mong muốn thực sự của con người trong các tình huống lý tưởng giả thuyết được quy định trong phương pháp CEV vẫn là bất khả, ngay cả với một siêu trí tuệ. Nếu không có thông tin về nội dung ý chí ngoại suy của chúng ta, AI sẽ không có được bất kỳ tiêu chuẩn đáng kể nào để hướng dẫn hành vi của nó. Tuy nhiên, mặc dù rất khó để biết chính xác những gì mà CEV của con người mong muốn, những phỏng đoán có căn cứ vẫn là hoàn toàn khả dĩ. Điều này là khả thi ngay cả vào thời điểm hiện tại, khi chúng ta chưa đạt đến siêu trí tuệ. Chẳng hạn, sẽ đáng tin cậy hơn khi CEV của chúng ta mong muốn có những người trong tương lai sẽ giàu có và hạnh phúc hơn so với ước nguyện rằng tất cả chúng ta ngồi trong một căn phòng tối và chịu đựng nỗi đau. Nếu chúng ta có thể đưa ra ít nhất là một vài phán đoán như vậy một cách hợp lý, thì một siêu trí tuệ cũng vậy. Ngay từ đầu, hành vi của siêu trí tuệ có thể được hướng dẫn bởi các ước lượng của nó về nội dung CEV của con người. Nó sẽ có lý do công cụ mạnh mẽ để tinh chỉnh các ước tính ban đầu này (chẳng hạn: bằng cách nghiên cứu văn hóa và tâm lý con người, quét bộ não con người và suy luận về cách hành xử của chúng ta nếu hiểu biết nhiều hơn, suy nghĩ rõ ràng hơn,...). Khi điều tra những vấn đề này, AI sẽ được các ước tính ban đầu về CEV của chúng ta hướng dẫn; sao cho, chẳng hạn, AI sẽ không cần chạy hàng loạt các bản mô phỏng với đầy sự đau khổ của con người nếu nó ước tính rằng CEV của chúng ta có thể sẽ quy kết những mô phỏng này là tội ác tâm trí.

Một ý kiến phản đối khác cho rằng trên thế giới có rất nhiều cách sống và quy tắc đạo đức khác nhau, đến mức việc trộn lẫn chúng vào một CEV có thể là bất khả. Ngay cả khi người ta có thể hòa trộn

chúng, thì kết quả cũng không hẳn sẽ “thơm ngon”: chúng ta khó mà có được một bữa ăn ngon bằng cách trộn tất cả các hương vị tốt nhất từ nhiều món ăn khoái khẩu của mọi người.¹³ Để đáp lại, chúng ta có thể chỉ ra rằng cách tiếp cận CEV không đòi hỏi tất cả các cách sống, quy tắc đạo đức hoặc giá trị cá nhân được hòa trộn với nhau thành một “món hầm”. Động lực CEV sẽ chỉ được hành động khi mong muốn của chúng ta đồng nhất với nhau. Về các vấn đề có sự bất đồng không thể hòa giải trên diện rộng, ngay cả sau khi các điều kiện lý tưởng hóa khác nhau đã được áp đặt, động lực này nên kiềm chế bản thân và không quyết định kết quả. Tiếp tục với hình ảnh nấu nướng, có thể các cá nhân hoặc nền văn hóa sẽ có những món ăn khoái khẩu khác nhau, nhưng họ vẫn sẽ đa phần đồng thuận rằng các món ăn không nên độc hại. Động lực CEV sau đó có thể hành động để ngăn chặn nguy cơ “ngộ độc thực phẩm” trong khi vẫn cho phép con người nấu nướng mà không có sự chỉ đạo hay can thiệp của nó.

Những lý do căn bản của CEV

Bài viết của Yudkowsky đã đưa ra bảy lập luận cho phương pháp CEV. Trong đó, ba lập luận về cơ bản là những cách khác nhau để đưa ra quan điểm cho rằng trong khi chúng ta nên đặt ra mục tiêu làm một việc gì đó nhân văn và hữu ích, sẽ rất khó để đưa ra một bộ quy tắc rõ ràng không có những diễn giải ngoài ý muốn và hậu quả không mong muốn.¹⁴ Cách tiếp cận CEV phải được áp dụng mạnh mẽ và tự tin; nó phải nắm bắt nguồn gốc các giá trị của chúng ta, thay vì dựa vào việc chúng ta liệt kê và xác định chính xác, một lần và mãi mãi, từng giá trị thiết yếu của mình.

Bốn ý còn lại là những lập luận sâu hơn (nhưng quan trọng), đưa ra những ước nguyện về các giải pháp ứng viên cho vấn đề đặc tả giá trị, và gợi ý rằng CEV đáp ứng được các ước nguyện này.

“Lý tưởng hóa môi trường tăng trưởng đạo đức”

Đây là ước nguyện đầu tiên, rằng giải pháp của chúng ta phải cân nhắc đến khả năng tiến bộ đạo đức. Như đã đề xuất trước đó, có nhiều lý do để tin rằng niềm tin đạo đức hiện tại của chúng ta vẫn còn nhiều thiếu sót; và có lẽ là những thiếu sót sâu sắc. Nếu quy định một quy tắc đạo đức cụ thể và không thể thay đổi để AI tuân theo, chúng ta sẽ bị khóa trong niềm tin đạo đức hiện tại (bao gồm cả những lỗi sai) và điều này sẽ phá hủy mọi hy vọng tăng trưởng đạo đức. Ngược lại, cách tiếp cận CEV sẽ tạo điều kiện cho quá trình tăng trưởng này vì AI đã cố gắng làm điều mà chúng ta mong muốn sẽ làm nếu được phát triển hơn nữa trong điều kiện thuận lợi, và có thể nếu phát triển như vậy, những niềm tin và sự nhạy bén về đạo đức của chúng ta sẽ loại trừ được những lỗ hổng và hạn chế của chính mình.

“Tránh cướp đoạt vận mệnh của loài người”

Yudkowsky đã suy nghĩ về một kịch bản mà một nhóm nhỏ lập trình viên tạo ra một AI hạt giống mà sau đó sẽ phát triển thành một siêu trí tuệ với ưu thế chiến lược quyết định. Trong kịch bản này, các lập trình viên ban đầu nắm trong tay toàn bộ tài nguyên vũ trụ của loài người. Rõ ràng, đây là một trách nhiệm quá nặng nề với bất kỳ người bình thường nào. Tuy nhiên, các lập trình viên không thể trốn tránh hoàn toàn trách nhiệm một khi họ thấy mình rơi vào tình huống này: bất kỳ lựa chọn nào họ đưa ra, kể cả từ bỏ dự án, sẽ đều có hậu quả mang tính lịch sử với thế giới. Yudkowsky coi CEV là một cách

để các lập trình viên tránh tự đeo vào cổ mình thứ đặc quyền (hay đúng hơn là gánh nặng) quyết định tương lai nhân loại. Bằng cách thiết lập một động lực thực thi ý chí ngoại suy kết hợp của loài người (trái với ý chí của chính họ, hoặc lý thuyết đạo đức ưa thích của riêng họ), họ thực sự phân phối ảnh hưởng của mình đến tương lai cho toàn nhân loại.

“Tránh tạo ra động lực cho con người hiện đại trong việc chống lại động lực ban đầu”

Phân phối ảnh hưởng trên tương lai của nhân loại không chỉ là ưu tiên về đạo đức của nhóm lập trình viên đang thực thi tầm nhìn ưa thích của mình, mà còn là cách để giảm động lực chiến đấu chống lại người/ tổ chức nào tạo ra siêu trí tuệ đầu tiên. Theo cách tiếp cận CEV, các lập trình viên (hoặc người tài trợ) không ảnh hưởng nhiều hơn đến nội dung của kết quả so với bất kỳ người nào khác, mặc dù họ hiển nhiên đóng vai trò quan trọng trong việc xác định cấu trúc của phép ngoại suy và việc loài người quyết định thực hiện CEV thay cho các phương pháp khác. Tránh xung đột là một việc quan trọng, không chỉ vì tác hại tức thời mà cuộc xung đột đó có xu hướng gây ra mà còn vì nó sẽ cản trở sự hợp tác trong quá trình đương đầu với thách thức khó khăn từ việc phát triển siêu trí tuệ một cách an toàn và có lợi.

CEV cần có khả năng kêu gọi sự hỗ trợ rộng rãi, không phải chỉ vì nó phân bổ ảnh hưởng một cách công bằng. Ngoài ra, lý do sâu xa hơn nằm ở tiềm năng mang lại hòa bình của CEV, cụ thể là nó cho phép nhiều nhóm khác nhau hy vọng rằng tầm nhìn ưa thích của họ về tương lai sẽ chiếm ưu thế hoàn toàn. Hãy tưởng tượng một thành viên của Taliban (từ Afghanistan) đang tranh luận với một thành viên

của Hiệp hội Nhân văn Thụy Điển. Hai thành viên này có thể giới quan rất khác nhau, và miền đất hứa của người này có thể là địa ngục với người kia. Cả hai người cũng không thể bị kích động bởi bất kỳ quan điểm thỏa hiệp nào, chẳng hạn như cho phép phụ nữ được hưởng sự giáo dục nhưng chỉ học đến lớp chín, hoặc cho phép phụ nữ Thụy Điển đi học nhưng phụ nữ Afghanistan thì không. Tuy nhiên, cả thành viên của Taliban và Hiệp hội Nhân văn đều có thể chứng thực một nguyên tắc, rằng tương lai nên được xác định bởi CEV của con người. Thành viên Taliban có thể lý giải rằng nếu quan điểm tôn giáo của anh ta thực sự đúng (như anh ta tin chắc là như vậy), và nếu có cơ sở vững chắc để chấp nhận những quan điểm này tồn tại (vì anh ta cũng bị thuyết phục), thì cuối cùng loài người sẽ chấp nhận những quan điểm này nếu họ ít thành kiến và thiên vị hơn, dành nhiều thời gian nghiên cứu kinh sách, có thể hiểu rõ hơn cách thế giới vận hành và nhận ra các ưu tiên thiết yếu, có thể được giải thoát khỏi sự nổi loạn phi lý và sự hèn nhát, cùng nhiều điều kiện khác. Loài người cũng sẽ tin rằng, dưới những điều kiện lý tưởng hóa này, họ sẽ đi theo những nguyên tắc vốn đã được tán thành.

“Duy trì trách nhiệm của loài người đối với vận mệnh của chính mình”

Chúng ta có thể không muốn một kết quả mà trong đó, một siêu trí tuệ có tính “gia trưởng” sẽ không ngừng theo dõi chúng ta, điều khiển các vấn đề của chúng ta ở tầm vi mô với mục tiêu tối ưu hóa mọi chi tiết theo một kế hoạch lớn. Ngay cả khi chúng ta quy định rằng siêu trí tuệ sẽ hoàn toàn nhân từ, cũng như không có tính tự phụ, kiêu ngạo, hống hách, hẹp hòi và những thiếu sót khác của con

người, chúng ta vẫn có thể nổi cơn thịnh nộ vì mất đi quyền tự chủ do sự sắp đặt như vậy. Chúng ta có thể thích tạo ra số phận của mình hơn, ngay cả khi điều đó có nghĩa là đôi khi chúng ta sẽ trở nên dò dẫm. Có lẽ chúng ta cũng muốn siêu trí tuệ đóng vai trò là mạng lưới an toàn, hỗ trợ mình khi mọi thứ trở nên tồi tệ, và để chúng ta tự bảo vệ mình khi không xảy ra điều đó.

CEV cho phép khả năng này xảy ra. CEV cần phải là một động lực ban đầu của người dùng, một quá trình chỉ chạy một lần rồi tự thay thế bản thân bằng bất cứ điều gì mà ý chí ngoại suy mong muốn. Nếu ý chí ngoại suy của loài người mong muốn chúng ta sống dưới sự giám sát của một AI “gia trưởng”, động lực CEV sẽ tạo ra một AI như vậy và trao cho nó dây cương. Còn nếu ý chí ngoại suy mong muốn một chính phủ dân chủ toàn cầu được thiết lập, động lực CEV có thể tạo điều kiện cho việc thành lập một tổ chức như vậy và vẫn giữ trạng thái “vô hình” trong các trường hợp khác. Thay vào đó, nếu ý chí ngoại suy của loài người muốn mỗi người nhận được một nguồn tài nguyên mà mình có thể sử dụng miễn là tôn trọng quyền bình đẳng của người khác, động lực CEV có thể biến điều này thành sự thật bằng cách hoạt động “ngầm” giống như một quy luật tự nhiên, nhằm ngăn chặn hành vi xâm phạm, trộm cắp, tấn công và các hành vi cưỡng chế khác.¹⁶

Do đó, cấu trúc của phương pháp CEV cho phép một phạm vi kết quả gần như không giới hạn. Chúng ta cũng có thể hình dung rằng ý chí ngoại suy của loài người muốn CEV không làm gì cả. Trong trường hợp đó, AI thực hiện CEV sẽ tự tắt sau khi đã xác minh xác suất phù hợp của việc ý chí ngoại suy muốn nó làm điều này.

Những nhận xét khác

Đề xuất về phương pháp CEV, như đã nêu, tất nhiên là giản đơn tốt nhất. Nó có một số tham số tự do có thể được chỉ định theo nhiều cách khác nhau, mang lại các phiên bản khác nhau.

Một tham số là cơ sở ngoại suy: Ý chí của những ai sẽ được đưa vào sử dụng? Chúng ta có thể nói là của tất cả mọi người, nhưng câu trả lời này sẽ làm phát sinh một loạt các câu hỏi khác. Liệu cơ sở ngoại suy có bao gồm những cá thể được gọi là “cá thể cận biên”, chẳng hạn như phôi thai, thai nhi, người chết não, bệnh nhân mắc chứng mất trí nhớ nghiêm trọng hoặc đang trong tình trạng sống thực vật vĩnh viễn? Có phải mỗi bán cầu não của một bệnh nhân não tách đôi (split-brain) đều có trọng số riêng trong ngoại suy, và liệu trọng số này có giống như toàn bộ não của một đối tượng bình thường? Còn những người sống trong quá khứ nhưng giờ đã chết, những người sẽ được sinh ra trong tương lai, các loài vật bậc cao hơn và các tạo vật có tri giác khác, tâm trí kỹ thuật số, hay người ngoài hành tinh thì sao?

Một lựa chọn sẽ chỉ bao gồm dân số người trưởng thành trên Trái đất còn sống khi bắt đầu sáng tạo AI. Phép ngoại suy ban đầu từ cơ sở này sau đó có thể quyết định liệu cơ sở đó có nên được mở rộng hay không. Do số lượng “cá thể cận biên” trong các khu vực ngoại vi của cơ sở này là khá nhỏ, kết quả của phép ngoại suy có thể không phụ thuộc nhiều vào chính xác nơi ranh giới được rút ra về việc nó có bao gồm cả (chẳng hạn như) thai nhi hay không.

Việc một cá thể nào đó bị loại khỏi cơ sở ngoại suy ban đầu không ngụ ý rằng mong muốn và hạnh phúc của cá thể này bị coi nhẹ. Nếu ý chí ngoại suy kết hợp của những cá thể trong cơ sở ngoại suy (ví dụ, con người trưởng thành còn sống) mong muốn sự cân

nhắc về đạo đức được mở rộng cho những sinh vật khác, kết quả của động lực CEV sẽ phản ánh sự ưu tiên này. Tuy nhiên, có thể lợi ích của những cá thể được đưa vào cơ sở ngoại suy ban đầu sẽ được mang lại ở một mức độ lớn hơn so với lợi ích của cá thể bên ngoài. Đặc biệt, nếu động lực chỉ hành động khi có sự thống nhất rộng rãi giữa các ý chí ngoại suy riêng lẻ (như trong đề xuất ban đầu của Yudkowsky), dường như sẽ có một rủi ro đáng kể về một cuộc “bỏ phiếu” chống có thể ngăn chặn việc bảo vệ, chẳng hạn như, lợi ích của các loài không phải con người hoặc tâm trí kỹ thuật số. Kết quả thu được có thể có khả năng suy thoái về đạo đức.¹⁷

Một động lực cho đề xuất CEV là tránh tạo động lực để con người chống lại việc xây dựng AI siêu trí tuệ đầu tiên. Mặc dù đề xuất CEV “ghi điểm” tốt hơn khi xét đến ước nguyện này so với nhiều phương pháp thay thế, nhưng nó không hoàn toàn loại bỏ các động lực phát sinh xung đột. Một cá nhân, nhóm hoặc quốc gia ích kỷ có thể tìm cách mở rộng tương lai của mình thông qua việc không cho những người khác gia nhập cơ sở ngoại suy.

Việc chiếm đoạt quyền lực như vậy có thể được hợp lý hóa theo nhiều cách khác nhau. Chẳng hạn, ai đó có thể lập luận rằng người tài trợ cấp vốn cho sự phát triển của AI xứng đáng sở hữu kết quả này. Đạo lý này có lẽ là sai. Chẳng hạn, ai đó cũng có thể phản đối rằng dự án tạo ra AI hạt giống thành công đầu tiên sẽ đặt ra một trạng thái rủi ro lớn từ bên ngoài đối với phần còn lại của nhân loại, và do đó có nghĩa vụ bù đắp. Phần bù đắp sẽ lớn đến mức chỉ có thể được hiện thực hóa thông qua việc cho mọi người sở hữu một phần lợi ích nếu mọi thứ đi vào ổn định.¹⁸

Một lập luận khác có thể được áp dụng để hợp lý hóa việc thu tóm quyền lực là phần lớn con người đều xuất phát từ những mong muốn xấu xa và việc đưa họ vào trong cơ sở ngoại suy sẽ tạo ra rủi ro khiến tương lai nhân loại trở thành địa ngục. Thật khó để biết được trong trái tim người bình thường có bao nhiêu phần tốt xấu, và cũng rất khó để biết sự cân bằng này khác nhau bao nhiêu giữa các nhóm, các tầng lớp xã hội, các nền văn hóa hay các quốc gia. Bất kể chúng ta có lạc quan hay bi quan về bản chất của con người, thì cũng sẽ không muốn đánh cược tài nguyên vũ trụ của con người với một sự suy đoán rằng với đa phần trong số bảy tỷ người hiện đang sống, phần thiện sẽ tồn tại trong ý chí ngoại suy. Tất nhiên, loại bỏ một nhóm người nhất định khỏi cơ sở ngoại suy không đảm bảo rằng “cái thiện” sẽ chiến thắng; và có thể tâm hồn của những người sẽ sớm loại bỏ người khác hay thu tóm quyền lực thường có xu hướng chứa đựng một khoảng tối lớn bất thường.

Tuy nhiên, một lý do khác để chống lại động lực ban đầu là người ta có thể tin rằng AI của một ai đó khác sẽ không hoạt động như quảng cáo, ngay cả khi AI này được rêu rao là một cách để thực hiện CEV của loài người. Nếu các nhóm khác nhau có niềm tin khác nhau về việc hệ thống nào có khả năng thành công nhất, họ có thể chiến đấu để ngăn chặn các nhóm khác khởi chạy. Sẽ tốt hơn trong những tình huống như vậy nếu các dự án cạnh tranh lẫn nhau có thể giải quyết sự khác biệt về nhận thức bằng một cách thức nào đó xác định người đúng chính xác hơn so với xung đột vũ trang.¹⁹

Các mô hình đạo đức

Đề xuất về CEV không phải là hình thức khả thể duy nhất của tính quy chuẩn gián tiếp. Ví dụ, thay vì thực thi ý chí ngoại suy kết

hợp của con người, người ta có thể cố gắng xây dựng một AI với mục tiêu làm điều đúng đắn về đạo đức, dựa vào năng lực nhận thức vượt trội của AI trong việc tìm ra hành động phù hợp với mô tả đó. Chúng ta có thể gọi đề xuất này là “tính đúng đắn về đạo đức” (moral rightness – MR). Ý tưởng ở đây là con người chúng ta có hiểu biết không hoàn hảo về điều gì là đúng và sai, và có lẽ còn kém hiểu biết hơn về cách phân tích triết học của MR. Tuy nhiên, một siêu trí tuệ lại có thể hiểu được những điều này cặn kẽ hơn.²⁰

Chúng ta sẽ làm gì nếu không chắc chắn về việc chủ nghĩa hiện thực đạo đức là đúng? Tất nhiên, chúng ta vẫn có thể thử áp dụng đề xuất về MR, và chỉ cần đảm bảo chỉ định cho AI những gì nên làm trong trường hợp rằng giả định của nó về chủ nghĩa hiện thực đạo đức là sai. Chẳng hạn, chúng ta có thể quy định nếu AI ước tính với xác suất phù hợp rằng không tồn tại sự thật phi tương đối phù hợp về tính đúng đắn về đạo đức, khi đó, nó nên đảo ngược quá trình và chuyển sang thực thi ý chí ngoại suy kết hợp, hoặc đơn giản hơn, nó nên tự tắt.²¹

MR xem ra sở hữu một số lợi thế so với CEV. MR sẽ loại bỏ các tham số tự do khác nhau trong CEV, chẳng hạn như mức độ kết hợp giữa các ý chí ngoại suy cần thiết để AI hành động theo kết quả, sự dễ dàng mà với đó đa số có thể đàn áp thiểu số bất đồng ý kiến, và bản chất của môi trường xã hội trong đó những bản thể ngoại suy của chính chúng ta được cho là đã cùng nhau trưởng thành hơn. MR xem ra sẽ loại bỏ khả năng xảy ra sai lỗi về đạo đức do sử dụng cơ sở ngoại suy quá hẹp hoặc quá rộng. Hơn nữa, MR sẽ định hướng để AI có hành động đúng đắn về đạo đức, ngay cả khi ý chí ngoại suy kết hợp của chúng ta vô tình mong muốn AI thực hiện những hành

động khả ố về đạo đức. Như đã lưu ý từ trước, đây có vẻ là một khả năng thực sự với đề xuất CEV. Sự tốt đẹp về đạo đức (moral goodness) có thể giống như một kim loại quý hơn là một yếu tố dư thừa trong bản chất con người, và ngay cả sau khi “quặng” này đã được xử lý và tinh chế theo các quy định của đề xuất về CEV, thì cũng chẳng ai biết được liệu kết quả chính yếu sẽ là đức hạnh sáng chói, một mớ gi sắt tầm thường, hay là một vũng sinh độc hại.

MR xem ra cũng có một số nhược điểm. Phương pháp tiếp cận này dựa trên khái niệm “đúng đắn về đạo đức” (morally right), một khái niệm cực kỳ khó hiểu mà các nhà triết học đã vật lộn từ thời cổ đại nhưng vẫn chưa đạt được sự đồng thuận trong phân tích. Lựa chọn một cách giải thích sai lầm về “tính đúng đắn về đạo đức” có thể dẫn đến những kết quả vô cùng sai lệch về đạo đức. Khó khăn này trong việc xác định “tính đúng đắn về đạo đức” dường như có thể khiến đề xuất về MR dễ bị sai lỗi hơn đáng kể. Tuy nhiên, việc đề xuất về MR thực sự là một bất lợi hữu hình trong vấn đề này vẫn chưa rõ ràng. Cả đề xuất về CEV cũng sử dụng các thuật ngữ và khái niệm khó giải thích (chẳng hạn như “hiểu biết”, “trở thành những hình mẫu mà mình mong muốn”, “đã cùng nhau trưởng thành hơn”, và nhiều thuật ngữ khác).²² Ngay cả khi những khái niệm này ít mơ hồ hơn so với “tính đúng đắn về đạo đức”, chúng vẫn quá xa vời so với tất cả những gì mà các lập trình viên hiện có thể thể hiện bằng mã.²³ Con đường dẫn đến việc trao cho AI bất kỳ khái niệm nào trong số này đều có thể liên quan đến việc cung cấp cho nó khả năng ngôn ngữ đại cương (ít nhất là có thể so sánh được với một người trưởng thành bình thường). Khả năng hiểu ngôn ngữ tự nhiên đại cương như vậy sau đó có thể được sử dụng để hiểu ý nghĩa của đạo đức.

Nếu AI nắm bắt được ý nghĩa, nó có thể tìm kiếm hành động phù hợp. Khi AI phát triển siêu trí tuệ, nó có thể đạt được tiến bộ trên hai mặt: vấn đề triết học trong việc hiểu được tính đúng đắn về đạo đức là gì, và vấn đề thực tế của việc áp dụng sự hiểu biết này vào việc đánh giá các hành động cụ thể.²⁴ Điều này sẽ không đơn giản, nhưng chúng ta cũng không biết rõ được liệu nó có khó hơn việc ngoại suy ý chí kết hợp của con người hay không.²⁵

Một vấn đề cơ bản hơn với MR là ngay cả khi khả thi, nó cũng có thể không cung cấp cho chúng ta những gì mình muốn hoặc những gì mình sẽ chọn nếu trở nên thông minh và hiểu biết hơn. Tất nhiên đây là tính năng thiết yếu của MR chứ không phải là một lỗi ngẫu nhiên. Tuy vậy, nó có thể là một tính năng cực kỳ có hại cho chúng ta.²⁶

Một ai đó có thể cố gắng bảo lưu ý tưởng cơ bản của mô hình MR trong khi giảm bớt sự đòi hỏi của nó bằng cách tập trung vào tính chấp nhận được về đạo đức: một ý tưởng cho rằng chúng ta có thể để AI theo đuổi CEV của con người, miễn là nó không có những hành động không chấp nhận được về đạo đức. Ví dụ: ta có thể hình thành mục tiêu sau đây cho AI:

Trong số các hành động chấp nhận được về đạo đức với AI, hãy thực hiện một hành động mà CEV của con người sẽ thích. Tuy nhiên, nếu một phần của hướng dẫn này không có ý nghĩa cụ thể, hoặc nếu chúng ta hoàn toàn bối rối về ý nghĩa của nó, hoặc nếu chủ nghĩa hiện thực đạo đức là sai, hoặc nếu hành động tạo ra một AI với mục tiêu này là không chấp nhận được về đạo đức, hãy tiến hành thao tác tự tắt có kiểm soát.²⁷

Hãy làm theo ý nghĩa mong muốn của hướng dẫn này.

Người ta vẫn có thể lo lắng rằng mô hình tính chấp nhận được về đạo đức (moral permissibility – MP) này thể hiện mức độ tôn trọng quá cao với các yêu cầu về đạo đức. Sự hy sinh mà nó kéo theo sẽ lớn đến mức nào tùy thuộc vào việc lý thuyết đạo đức nào là đúng.²⁸ Nếu đạo đức có tính thỏa mãn vừa đủ, nghĩa là coi những hành động tuân thủ một vài điều hạn chế đạo đức cơ bản là chấp nhận được về đạo đức, MP có thể nhường chỗ để ý chí ngoại suy kết hợp của chúng ta tác động đến hành động của AI. Tuy nhiên, nếu đạo đức có tính tối đa hóa, chẳng hạn, nếu hành động chấp nhận được về đạo đức chỉ là những hành động có hệ quả đạo đức tốt nhất, MP có thể không nhường chỗ hoặc dành rất ít không gian cho những sở thích của chúng ta định hình kết quả.

Để minh họa vấn đề này, hãy cùng dành ít phút để quay trở lại với ví dụ về hệ quả luận theo chủ nghĩa khoái lạc. Giả sử lý thuyết đạo đức này là đúng và AI biết điều này. Đối với các mục đích hiện tại, chúng ta có thể định nghĩa hệ quả luận theo chủ nghĩa khoái lạc là tuyên bố cho rằng một hành động là đúng đắn về đạo đức (và được phép về đạo đức) nếu và chỉ nếu trong số tất cả các hành động khả thi, không có hành động nào khác tạo ra được một sự cân bằng lớn hơn của tỷ lệ khoái lạc trên đau khổ. AI, theo MP, có thể tối đa hóa niềm vui bằng cách chuyển đổi phần vũ trụ có thể tiếp cận thành hedonium, một quá trình có thể liên quan đến việc xây dựng computronium, và sử dụng nó để thực hiện các tính toán tạo ra những trải nghiệm thú vị. Vì mô phỏng bất kỳ bộ não con người nào hiện có không phải là cách tạo ra khoái cảm hiệu quả nhất, nên hậu quả có thể xảy ra là tất cả chúng ta đều sẽ diệt vong.

Bằng cách thực thi đề xuất về MR hoặc MP, chúng ta sẽ có nguy cơ hy sinh mạng sống của mình vì một điều tốt đẹp hơn. Đây sẽ là một sự hy sinh lớn hơn bất kỳ ai có thể nghĩ, bởi vì những gì chúng ta sẽ mất không chỉ là cơ hội để sống một cuộc sống bình thường của con người, mà còn là cơ hội để tận hưởng cuộc sống lâu dài và giàu có hơn mà một siêu trí tuệ thân thiện có thể ban tặng.

Sự hy sinh xem ra còn kém hấp dẫn hơn khi chúng ta ngẫm lại rằng siêu trí tuệ có thể nhận ra một điều tốt đẹp gần tương đương (về mặt tỷ lệ) trong khi hy sinh một định lượng ít hơn rất nhiều trong sự hạnh phúc có thể có của chúng ta. Giả sử chúng ta đã đồng ý cho phép gần như toàn bộ phần vũ trụ có thể tiếp cận được chuyển đổi thành hedonium, tất cả mọi thứ trừ một phần bảo lưu nhỏ, chẳng hạn như Dải Ngân hà, được đặt sang một bên để đáp ứng nhu cầu của chúng ta. Sau đó, vẫn sẽ có 100 tỷ thiên hà dành cho việc tối đa hóa khoái lạc, nhưng chúng ta sẽ có một thiên hà để tạo ra những nền văn minh tuyệt vời tồn tại đến hàng tỷ năm, và trong đó, con người và động vật có thể tồn tại và thịnh vượng, đồng thời có cơ hội phát triển thành những thực thể tinh thần hậu nhân (posthuman).²⁹

Việc một ai đó thích lựa chọn thứ hai này (giống như tôi) hàm chỉ người đó không có sở thích chi phối về mặt ngôn từ vô điều kiện trong việc thực hiện hành động chấp nhận được về đạo đức. Nhưng nó phù hợp với việc đặt trọng tâm vào vấn đề đạo đức.

Ngay cả từ quan điểm đạo đức thuần túy, có thể sẽ tốt hơn khi chúng ta ủng hộ một số đề xuất ít chú trọng về đạo đức hơn MR hoặc MP. Nếu đề xuất lý tưởng nhất về đạo đức không có cơ hội được thực thi, có thể vì tính đòi hỏi của nó, sẽ tốt hơn về đạo đức nếu chúng ta đưa ra một đề xuất nào đó khác gần như lý tưởng và sở

hữu cơ hội được thực thi gia tăng đáng kể khi được chúng ta thúc đẩy.³⁰

“Hãy làm điều tôi muốn”

Chúng ta có thể cảm thấy không chắc chắn về việc nên chọn CEV, MR, MP, hay một phương pháp tiếp cận khác. Liệu chúng ta cũng có thể đặt cược vào quyết định cấp cao hơn đang nói đến, nghĩa là truyền tải nhiều hơn nữa công việc nhận thức cho AI hay không? Đây là giới hạn cho sự lười biếng mà chúng ta có thể có?

Chẳng hạn, hãy xem xét mục tiêu “dựa trên lý do” sau đây:

Hãy làm bất kỳ điều gì mà chúng ta có lý do hợp lý nhất để yêu cầu AI thực hiện.

Mục tiêu này có thể được cụ thể hóa thành ý chí ngoại suy hoặc đạo đức, hoặc một thứ khác, nhưng dường như nó sẽ giải thoát chúng ta khỏi những nỗ lực và rủi ro liên quan đến việc cố gắng chọn ra mục tiêu mà chúng ta có lý do hợp lý nhất để lựa chọn trong số các mục tiêu cụ thể hơn này.

Tuy nhiên, ở đây cũng có một số vấn đề với các mục tiêu dựa trên đạo đức. Thứ nhất, chúng ta có thể sợ rằng những mong muốn của mình có quá ít “đất diễn” với mục tiêu dựa trên lý do này. Một số triết gia cho rằng một cá nhân luôn có lý do để làm những gì tốt nhất về đạo đức cho mình. Nếu các triết gia đó đúng, mục tiêu dựa trên lý do này sẽ thiên về phía đề xuất MR, và đồng thời tạo ra rủi ro của việc một siêu trí tuệ thực hiện một động lực như vậy sẽ giết chết bất kỳ ai nó có thể tiếp cận. Thứ hai, với việc tất cả các đề xuất đều được diễn tả thông qua ngôn ngữ kỹ thuật, chúng ta rất có khả năng đã hiểu sai ý nghĩa của những xác nhận mà chính mình đưa ra. Chúng ta đã thấy, đối với các mục tiêu dựa trên đạo đức, việc yêu cầu AI

làm điều đúng đắn có thể dẫn đến những hậu quả khôn lường ngoài ý muốn mà nếu đoán được, chúng ta sẽ không thực hiện mục tiêu. Điều tương tự cũng đúng với việc yêu cầu AI làm những gì chúng ta có lý do hợp lý nhất để làm.

Sẽ ra sao nếu chúng ta cố lảng tránh những khó khăn này bằng cách mô tả một mục tiêu với ngôn từ hoàn toàn phi kỹ thuật, chẳng hạn như “sự tử tế”:³¹

Hãy thực hiện hành động tử tế nhất, hoặc nếu không có hành động nào là tử tế nhất, hãy thực hiện một hành động “siêu” tử tế.

Sẽ khó có lý do nào phù hợp để phản đối việc tạo ra một AI tử tế. Tuy nhiên, chúng ta cũng cần xem xét xem ý nghĩa chính xác của diễn đạt này. Kho từ vựng của chúng ta có rất nhiều ý nghĩa khác nhau của phạm trù “tử tế” không được sử dụng ở đây: chúng ta không muốn AI lễ phép và lịch sự, hay kén chọn và quá tinh tế. Nếu chúng ta có thể tin tưởng vào việc AI nhận ra cách lý giải mong muốn về “sự tử tế” và có động lực để theo đuổi cách lý giải đó, mục tiêu này dường như sẽ tương đương một mệnh đề AI làm những gì mà lập trình viên nhắm tới.³² Một lệnh có hiệu quả tương tự đã được đưa vào trong công thức của CEV (“... giải thích như cách ta muốn nó được giải thích”) và trong tiêu chí về tính chấp nhận được về đạo đức (“Hãy làm theo ý nghĩa mong muốn của hướng dẫn này”). Bằng cách gán một mệnh đề kiểu “hãy làm điều tôi muốn” như vậy, chúng ta có thể chỉ ra rằng các từ khác trong phần mô tả mục tiêu nên được hiểu theo nhu cầu chứ không theo nghĩa đen. Nhưng nói rằng AI nên hành xử “tử tế” lại gần như không bổ sung được điều gì: công việc thực sự chỉ được thực hiện theo hướng dẫn “hãy làm điều tôi muốn”. Nếu biết cách mã hóa mệnh đề “hãy làm điều tôi muốn” một cách phổ

quát và mạnh mẽ, chúng ta cũng có thể sử dụng nó như một mục tiêu độc lập.

Chúng ta có thể thực thi một động lực kiểu “hãy làm gì tôi muốn” như thế nào? Nghĩa là, làm thế nào tạo ra được một AI có động lực để diễn giải một cách hào phóng những mong muốn và ý định hàm ẩn rồi theo đó hành động? Bước đi ban đầu có thể là cố gắng làm rõ hơn những gì chúng ta cần khi ra lệnh “hãy làm điều tôi muốn”. Sẽ hữu ích nếu chúng ta diễn giải cụ thể mệnh lệnh này về mặt hành vi, chẳng hạn như về các sở thích được tiết lộ trong các tình huống giả thuyết khác nhau, chẳng hạn như khi chúng ta có nhiều thời gian hơn để xem xét các lựa chọn, thông minh hơn, biết nhiều hơn về các sự kiện liên quan, hay theo một cách nào đó khác, điều kiện sẽ thuận lợi hơn cho chúng ta thể hiện chính xác những gì mình muốn về một AI thân thiện, hữu ích, tử tế,... qua những lựa chọn cụ thể.

Ở đây, hiển nhiên chúng ta đã đi thành một vòng tròn và quay trở lại điểm xuất phát, đó là cách tiếp cận CEV cho quy chuẩn gián tiếp. Cách tiếp cận này về bản chất đã loại bỏ tất cả nội dung cụ thể khỏi phần đặc tả giá trị, chỉ để lại một giá trị trừu tượng được xác định đơn thuần về mặt quy trình: làm điều mà chúng ta muốn AI làm trong những tình huống được lý tưởng hóa phù hợp. Bằng cách quy chuẩn gián tiếp như vậy, chúng ta có thể hy vọng truyền sang AI phần lớn công việc nhận thức mà chính chúng ta sẽ cố gắng thực hiện nếu muốn đưa ra một mô tả cụ thể hơn về những giá trị mà AI sẽ theo đuổi. Vì thế, khi tìm cách tận dụng tối đa ưu thế tri thức của AI, CEV có thể được coi là một ứng dụng của nguyên tắc tôn trọng nhận thức.

Danh mục thành phần

Đến giờ, chúng ta đã xem xét các lựa chọn khác nhau cho việc nội dung nào sẽ được đưa vào hệ thống mục tiêu. Tuy nhiên, hành vi của AI cũng sẽ bị ảnh hưởng bởi các lựa chọn thiết kế khác. Cụ thể, loại lý thuyết quyết định và nhận thức luận mà nó sử dụng có thể dẫn đến một sự khác biệt trọng yếu. Một câu hỏi quan trọng khác là liệu các kế hoạch của AI có được con người xem xét trước khi thực hiện hay không.

Bảng 13 Danh mục thành phần

Nội dung mục tiêu	AI nên theo đuổi mục tiêu nào? Bản mô tả mục tiêu nên được hiểu ra sao? Mục tiêu có nên bao gồm việc trao phần thưởng đặc biệt cho những người đã đóng góp cho thành công của dự án?
Lý thuyết quyết định	AI nên sử dụng lý thuyết quyết định nhân quả, lý thuyết quyết định theo bằng chứng, lý thuyết quyết định phi cập nhật, một lý thuyết khác?
Nhận thức luận	AI nên có hàm xác suất tiên nghiệm nào và nên đưa ra những giả định cụ thể hay hàm chỉ nào về thế giới? Nó nên sử dụng lý thuyết vị nhân (anthropics) nào?
Quá trình phê chuẩn	Các kế hoạch của AI có nên được xem xét bởi con người trước khi đưa vào hoạt động hay không? Nếu có, giao thức cho quá trình xem xét đó là gì?

Bảng 13 tóm tắt các lựa chọn thiết kế này. Một dự án xây dựng siêu trí tuệ phải giải thích được những lựa chọn đã đưa ra liên quan đến từng thành phần này cũng như lý giải nguyên do của những lựa chọn đó.³³

Nội dung mục tiêu

Chúng ta đã thảo luận về cách thức mà chuẩn hóa gián tiếp có thể được sử dụng trong việc chỉ định các giá trị mà AI sẽ theo đuổi, đồng thời cũng đã thảo luận về một số lựa chọn, như các mô hình dựa trên đạo đức và ý chí ngoại suy kết hợp. Với mỗi lựa chọn như vậy, chúng ta sẽ cần đưa ra thêm các lựa chọn khác. Chẳng hạn, phương pháp CEV có nhiều biến thể, tùy thuộc vào người được đưa vào cơ sở ngoại suy, cấu trúc của phép ngoại suy, và nhiều yếu tố khác. Các hình thức lựa chọn động lực khác có thể đòi hỏi các loại nội dung mục tiêu khác nhau. Ví dụ, một hệ thống tiên tri có thể được xây dựng để đặt giá trị vào mục tiêu đưa ra câu trả lời chính xác. Hệ thống tiên tri được tạo ra với động lực tự hạn định cũng có thể có nội dung mục tiêu không đặt giá trị vào việc sử dụng quá mức tài nguyên nhằm đưa ra câu trả lời.

Một lựa chọn thiết kế khác cần được thực hiện giữa việc có và không đưa các điều khoản đặc biệt vào nội dung mục tiêu nhằm trao thưởng cho các cá nhân đóng góp vào việc hiện thực hóa thành công AI, như cho họ thêm tài nguyên hoặc tầm ảnh hưởng lên hành vi của AI. Chúng ta có thể gọi các điều khoản như vậy là các “gói phần thưởng” (incentive wrapping). Gói phần thưởng có thể được coi là một cách để tăng khả năng thành công của dự án, đổi lấy việc mục tiêu mà dự án đặt ra sẽ bị xâm phạm ở một mức độ nào đó.

Ví dụ, nếu dự án đặt mục tiêu tạo ra một động lực thực hiện ý chí ngoại suy kết hợp của con người, một gói phần thưởng có thể chỉ định rằng ý chí của một số cá nhân cụ thể cần được ưu tiên hơn trong phép ngoại suy. Nếu dự án như vậy thành công, kết quả sẽ không nhất thiết là việc thực thi ý chí ngoại suy kết hợp của loài

người. Thay vào đó, một mục tiêu gần xấp xỉ như vậy sẽ được hoàn tất.³⁴

Vì gói phần thưởng sẽ là một phần của nội dung mục tiêu mà siêu trí tuệ cần hiểu và theo đuổi, nó có thể lợi dụng quy chuẩn gián tiếp để xác định các điều khoản tinh vi, phức tạp mà người quản lý khó thực hiện. Chẳng hạn, thay vì thưởng cho các lập trình viên theo một chỉ tiêu thô nhưng dễ tiếp cận, như số giờ làm việc hay số lỗi sai đã khắc phục, gói phần thưởng có thể chỉ định rằng các lập trình viên “phải có phần thưởng tỷ lệ thuận với mức độ gia tăng mà những đóng góp của họ đã mang lại cho xác suất thành công dự kiến của dự án như người tài trợ mong muốn. Hơn nữa, sẽ không có lý do để giới hạn gói phần thưởng trong phạm vi cán bộ dự án. Thay vào đó, nó có thể chỉ định rằng mọi người đều cần được thưởng cho công lao của mình. Phân bổ công trạng là một vấn đề khó khăn, nhưng chúng ta có thể trông đợi siêu trí tuệ tiệm cận hợp lý các tiêu chí do gói phần thưởng đưa ra, một cách cụ thể hoặc hàm ẩn.

Chúng ta có thể hình dung được việc siêu trí tuệ thậm chí có thể tìm ra cách nào đó để thưởng cho những cá nhân đã qua đời trước khi nó sinh ra.³⁵ Gói phần thưởng sau đó có thể được mở rộng để áp dụng cho ít nhất là một phần trong số những người đã chết, có thể bao gồm cả những người chết trước khi dự án được thai nghén hay thậm chí trước cả khi khái niệm về gói phần thưởng lần đầu tiên được đưa ra. Tuy thể chế của một chính sách hồi tố như vậy sẽ không trao thưởng cho những ai đã “xanh cỏ” khi cuốn sách này được viết, nhưng nó có thể được ủng hộ vì lý do đạo đức – mặc dù có thể cho rằng chừng nào sự công bằng còn là mục tiêu thì nó nên

được đưa vào như một phần của nội dung đặc tả mục tiêu chứ không phải của gói phần thưởng.

Ở đây, chúng ta không thể đi sâu vào tất cả các vấn đề đạo đức và chiến lược liên quan đến gói phần thưởng. Tuy nhiên, quan điểm của một dự án đối với những vấn đề này sẽ là một khía cạnh quan trọng trong ý tưởng thiết kế cơ bản của dự án đó.

Lý thuyết quyết định

Một lựa chọn thiết kế quan trọng khác là lý thuyết quyết định mà AI nên được xây dựng để sử dụng. Điều này có thể ảnh hưởng đến cách hành xử của AI trong những tình huống quyết định về chiến lược. Chẳng hạn, lý thuyết quyết định có thể xác định liệu AI có dễ dàng giao dịch với (hoặc bị chiếm đoạt lợi ích bởi) các nền văn minh siêu trí tuệ khác tồn tại theo giả thuyết hay không. Các chi tiết của lý thuyết quyết định có thể cũng có ý nghĩa trong các tình huống liên quan đến các xác suất hữu hạn của các khoản thưởng phạt vô hạn (thuyết đặt cược của Pascal) hoặc các xác suất cực kỳ nhỏ của các khoản thưởng phạt hữu hạn cực lớn (thuyết cướp đoạt theo kiểu Pascal), hoặc trong bối cảnh AI đối mặt với sự không chắc chắn quy chuẩn cơ bản hay khi nhiều bản sao của một chương trình tác tử cùng tồn tại.³⁶

Các lựa chọn sẵn có bao gồm lý thuyết quyết định nhân quả (với nhiều biến thể) và lý thuyết quyết định theo bằng chứng. Bên cạnh đó, ta có các “ứng viên” mới hơn như lý thuyết quyết định phi thời gian (timeless) và lý thuyết quyết định phi cập nhật (updateless), vốn vẫn đang được phát triển.³⁷ Việc xác định và áp dụng thuyết quyết định phù hợp, cũng như giữ vững niềm tin rằng chúng ta đã hiểu đúng nó, có thể tương đối khó khăn. Tuy triển vọng xác định trực tiếp

một lý thuyết quyết định của AI có lẽ có nhiều hy vọng hơn so với việc xác định trực tiếp các giá trị cuối cùng của nó, nhưng chúng ta vẫn phải đối mặt với một nguy cơ sai lỗi đáng kể. Nhiều “biến chứng” có thể phá vỡ các lý thuyết quyết định phổ biến nhất hiện nay cũng chỉ vừa được tìm thấy, và điều này cho thấy có thể còn tồn tại những vấn đề khác chưa lộ diện. Kết quả của việc đưa ra cho AI một lý thuyết quyết định sai có thể gây ra một thảm họa sống còn.

Trước những khó khăn này, chúng ta có thể cân nhắc một cách tiếp cận gián tiếp để chỉ định lý thuyết quyết định mà AI nên sử dụng. Cách thức chính xác để thực hiện điều đó vẫn chưa rõ ràng. Chúng ta có thể muốn AI sử dụng “chính lý thuyết quyết định D mà chúng ta sẽ muốn nó sử dụng nếu đã suy nghĩ kỹ”. Tuy nhiên, AI sẽ cần có khả năng đưa ra quyết định trước khi tìm hiểu D là gì. Vì thế, ta cần một lý thuyết quyết định tạm thời hiệu quả nào đó (tạm gọi là D') để chi phối quá trình tìm hiểu về D của AI. Ai đó có thể cố gắng định nghĩa D' như là một sự chòng chập các giả thuyết hiện tại của AI về D (với trọng số dựa trên xác suất), mặc dù một số vấn đề kỹ thuật về việc làm thế nào để thực hiện điều này một cách hoàn toàn tổng quát vẫn chưa được giải quyết.³⁸ Cũng có lý do để lo ngại rằng AI có thể đưa ra những quyết định vô cùng tồi tệ (như tự lập trình lại để chạy theo lý thuyết quyết định sai) trong giai đoạn học tập trước khi có cơ hội để xác định lý thuyết quyết định cụ thể nào là đúng. Để giảm nguy cơ “trật bánh” trong giai đoạn dễ bị tổn hại này, thay vào đó, chúng ta có thể cố gắng cung cấp cho AI hạt giống một hình thức nào đó của tính hợp lý hạn chế (restricted rationality): một lý thuyết quyết định đơn giản nhưng đáng tin cậy bỏ qua những cân nhắc đặc thù, kể cả những gì mà chúng ta cho là hoàn toàn hợp lệ, và được thiết kế để

thay thế chính nó bằng một lý thuyết quyết định phức tạp hơn (được chỉ định gián tiếp) khi một số điều kiện nhất định được đáp ứng. Tính khả thi và cách thức thực hiện điều này là một câu hỏi nghiên cứu còn để ngỏ.

Nhận thức luận

Một dự án cũng sẽ cần đưa ra lựa chọn thiết kế cơ bản cho nhận thức luận của AI và chỉ định các nguyên tắc và tiêu chí mà theo đó, các giả thuyết thực nghiệm sẽ được đánh giá. Trong khung Bayes, chúng ta có thể nghĩ đến nhận thức luận như một hàm xác suất tiên nghiệm, hay việc AI gán xác suất cho các thế giới khả thể trước khi xem xét bất kỳ bằng chứng tri giác nào. Trong các khung khác, nhận thức luận có thể có dạng thức khác; nhưng một quy tắc học quy nạp nào đó sẽ là cần thiết trong mọi trường hợp, nếu AI phải khái quát hóa từ các quan sát trong quá khứ và dự đoán về tương lai.⁴⁰ Tuy nhiên, như với nội dung mục tiêu và lý thuyết quyết định, phần đặc tả của chúng ta về nhận thức luận cũng có nguy cơ không đạt kết quả mong muốn.

Ai đó có thể nghĩ rằng có một giới hạn về mức độ thiệt hại khả phát từ nhận thức luận được chỉ định không chính xác. Nếu nhận thức luận sai lỗi quá mức, AI sẽ không thể đạt mức rất thông minh và cũng không thể gây ra loại hình rủi ro được thảo luận trong cuốn sách này. Tuy nhiên, quan ngại nằm ở chỗ chúng ta có thể chỉ định một nhận thức luận đủ tốt để khiến AI có được hiệu quả mang tính công cụ trong phần lớn các tình huống, nhưng cũng có sai lỗi khiến nó đi chệch hướng trong một vấn đề trọng yếu nào đó. Một AI như vậy có thể giống với một người nhanh trí sở hữu thế giới quan được xác định dựa trên một giáo điều sai lầm và bám vào đó với một niềm

tin tuyệt đối, và cuối cùng sẽ “đánh nhau với cối xay gió” và toàn tâm theo đuổi những mục tiêu ảo tưởng và nguy hại.

Một số khác biệt nhỏ trong tiên nghiệm của AI có thể dẫn đến những sai khác lớn trong hành xử. Ví dụ, một AI có thể nhận được tiên nghiệm gán xác suất bằng không cho việc vũ trụ là vô hạn. Cho dù có bao nhiêu bằng chứng thiên văn mà nó tích lũy chứng minh điều ngược lại, một AI như vậy vẫn sẽ ngoan cố từ chối bất kỳ lý thuyết vũ trụ nào hàm chỉ một vũ trụ vô tận; kết quả là nó có thể đưa ra những lựa chọn dại dột.⁴¹ Hoặc, một AI có thể nhận được tiên nghiệm đưa ra xác suất bằng không cho việc vũ trụ không thể được tính toán theo kiểu Turing (thực tế đây là một đặc điểm chung của nhiều tiên nghiệm được thảo luận trong tài liệu này, bao gồm cả độ phức tạp Kolmogorov được đề cập trong Chương 1), một lần nữa với những hậu quả được hiểu sơ sài nếu giả định được đặt ra – “luận đề Church-Turing” – là sai. Một AI cũng có thể có tiên nghiệm đưa ra những ràng buộc siêu hình mạnh mẽ bằng cách này hay cách khác, ví dụ bằng cách loại trừ xác suất tiên nghiệm của việc bất kỳ dạng thức mạnh nào của nhị nguyên luận tâm-thân là đúng, hoặc xác suất của việc tồn tại các sự thật đạo đức bất biến. Nếu bất kỳ ràng buộc nào trong số đó bị nhầm lẫn, AI có thể tìm cách hiện thực hóa các mục tiêu tối thượng của nó theo những cách mà chúng ta coi là hiện thực hóa sai lệch. Tuy nhiên, không có lý do rõ ràng cho việc một AI như vậy, tuy mắc sai lầm cơ bản về một vấn đề quan trọng, lại không thể có đủ hiệu quả mang tính công cụ để đảm bảo ưu thế chiến lược quyết định. (Vị nhân học, ngành nghiên cứu về cách suy luận từ thông tin mang tính chỉ biệt (indexical information) với sự hiện diện của các hiệu ứng công suất tính toán, là một lĩnh vực khác trong đó

việc lựa chọn các tiên đề nhận thức luận có thể chứng tỏ tính then chốt.⁴²⁾

Chúng ta có thể nghi ngờ một cách hợp lý khả năng của mình trong việc kịp thời giải quyết tất cả các vấn đề cơ bản trong nhận thức luận để xây dựng AI hạt giống đầu tiên. Do đó, chúng ta có thể xem xét thực hiện một cách tiếp cận gián tiếp để chỉ định nhận thức luận của AI. Điều này sẽ làm phát sinh nhiều vấn đề giống như khi thực hiện một cách tiếp cận gián tiếp để xác định lý thuyết quyết định của nó. Tuy nhiên, với nhận thức luận, có thể có nhiều hy vọng về sự hội tụ “lành tính”, với bất kỳ dạng nhận thức luận nào cung cấp nền tảng đầy đủ cho một AI an toàn và hiệu quả, và cuối cùng sẽ mang lại các kết quả đáng tin cậy. Lý do cho điều này nằm ở chỗ bằng chứng và phân tích thực nghiệm đủ phong phú sẽ có xu hướng loại bỏ bất kỳ sự khác biệt trung bình nào trong các kỳ vọng tiên nghiệm.⁴³⁾

Một mục tiêu tốt sẽ là việc giao cho AI các nguyên tắc nhận thức luận cơ bản phù hợp với những gì chi phối suy nghĩ của chúng ta. Bất kỳ AI nào đi chệch khỏi nguyên tắc lý tưởng này đều sẽ được đánh giá là lập luận không đúng đắn nếu chúng ta áp dụng nhất quán các tiêu chuẩn của mình. Điều này hiển nhiên chỉ áp dụng cho các nguyên tắc nhận thức luận cơ bản của chúng ta. Các nguyên tắc không cơ bản cần được liên tục tạo ra và sửa đổi bởi chính AI hạt giống trong khi phát triển hiểu biết về thế giới. Điểm mấu chốt của siêu trí tuệ nằm ở việc loại trừ sự ngờ nghệch của chúng ta chứ không phải “nối giáo” cho những định kiến của con người.

Quá trình phê chuẩn

Mục cuối cùng trong danh sách lựa chọn thiết kế của chúng tôi là phê chuẩn (ratification). Các kế hoạch của AI có nên được con người

xem xét trước khi đưa vào áp dụng không? Đối với một AI tiên tri, câu hỏi này được trả lời ngầm trong sự khẳng định. AI tiên tri sẽ đưa ra thông tin; người đánh giá sẽ lựa chọn xem liệu và làm thế nào để hành động theo thông tin đó. Tuy nhiên, đối với AI “thần đèn” (genies), AI toàn năng (sovereign) và AI công cụ (tool), câu hỏi về việc có nên sử dụng một số hình thức phê chuẩn hay không vẫn còn bỏ ngỏ.

Để minh họa cách thức mà quá trình phê chuẩn có thể hoạt động, hãy xem xét một AI có mục đích hoạt động như một hệ thống toàn năng thực thi CEV của con người. Thay vì khởi động trực tiếp AI này, hãy tưởng tượng rằng trước tiên chúng ta đã xây dựng một AI tiên tri với mục đích duy nhất là trả lời các câu hỏi về những gì AI toàn năng sẽ làm. Như đã hé lộ trong các chương trước, có những rủi ro trong việc tạo ra một AI tiên tri siêu trí tuệ (chẳng hạn như nguy cơ gây ra tội ác tâm trí hoặc dư thừa hạ tầng). Nhưng với mục đích nêu trong ví dụ này, hãy giả sử rằng AI tiên tri đã được xây dựng thành công và tránh được những cạm bẫy đó.

Khi đó, chúng ta có một AI tiên tri cung cấp những dự đoán tốt nhất về hậu quả của việc chạy một số đoạn mã với mục đích thực thi CEV của loài người. AI tiên tri có thể không đoán được chi tiết những gì sẽ xảy ra, nhưng dự đoán của nó có thể sẽ tốt hơn so với chúng ta. (Nếu ngay cả một siêu trí tuệ cũng không thể dự đoán bất cứ điều gì mà đoạn mã sẽ làm, hẳn chúng ta sẽ phát điên khi chạy nó.) Vì vậy, AI tiên tri sẽ cần suy nghĩ một lúc rồi mới đưa ra dự đoán. Để khiến câu trả lời trở nên dễ hiểu, AI tiên tri có thể cung cấp cho người vận hành một loạt công cụ để khám phá các tính chất khác nhau của kết quả dự đoán. AI tiên tri có thể hiển thị hình ảnh về tương lai và

cung cấp số liệu thống kê về số lượng cá thể có tri giác sẽ tồn tại ở các thời điểm khác nhau, cùng với mức độ hạnh phúc trung bình, cao nhất và thấp nhất. Nó có thể cung cấp tiểu sử “kín” của một số cá thể được chọn ngẫu nhiên (có thể là những người tưởng tượng được chọn là đại diện). Nó có thể làm nổi bật các khía cạnh của tương lai mà nhà điều hành chưa nghĩ đến việc tìm hiểu nhưng sẽ được coi là thích hợp khi được chỉ ra.

Khả năng xem trước kết quả như vậy có lợi thế rõ ràng. “Bản xem trước” này có thể tiết lộ hậu quả của một lỗi trong thông số thiết kế hoặc mã nguồn của AI toàn năng dự kiến. Nếu “quả cầu pha lê” này cho thấy một tương lai bị hủy hoại, chúng ta có thể loại bỏ mã cho AI toàn năng dự kiến và thử làm một thứ khác. Có lý do để nói rằng chúng ta nên tự làm quen với các “phân nhánh” cụ thể của một lựa chọn trước khi cam kết với nó, đặc biệt là khi toàn bộ tương lai đang bị đe dọa.

Điều có lẽ ít rõ ràng hơn là phê chuẩn cũng có những nhược điểm đáng kể. Phẩm chất “ôn hòa” của CEV có thể bị suy yếu nếu các phe phái đối lập, thay vì yêu cầu một trí tuệ siêu việt phán xử với kỳ vọng được minh oan, có thể thấy trước phán quyết sẽ ra sao. Người ủng hộ cách tiếp cận dựa trên đạo đức có thể lo lắng rằng quyết tâm của người tài trợ sẽ sụp đổ nếu tất cả những hy sinh cần có để tối ưu hóa về đạo đức được tiết lộ. Tất cả chúng ta có thể có lý do để ưu ái một tương lai chứa đựng ít nhiều bất ngờ, một số bất đồng, một chút hoang dại, và một số cơ hội tự vượt qua chính mình – một tương lai với ranh giới không quá phù hợp để đưa ra các định kiến, nhưng lại có chỗ cho sự vận động mạnh mẽ và tăng trưởng ngoài kế hoạch. Chúng ta sẽ có ít khả năng sở hữu một góc nhìn mở

rộng như vậy nếu có thể chọn lọc từng chi tiết của tương lai và gửi trả lại bất kỳ “bản nháp” nào không hoàn toàn phù hợp với ý muốn của mình lúc đó.

Vì vậy, vấn đề phê chuẩn người tài trợ là không rõ ràng như những gì chúng ta thấy ban đầu. Tuy nhiên, sau khi cân nhắc, việc tận dụng cơ hội có được một bản xem trước như vậy xem ra sẽ là khôn ngoan nếu chức năng này có sẵn. Nhưng thay vì để người đánh giá tinh chỉnh mọi khía cạnh của kết quả, chúng ta có thể cho họ một quyền phủ quyết đơn giản chỉ có thể thực thi một vài lần trước khi toàn bộ dự án bị bãi bỏ.⁴⁴

Tiến tới đủ gần

Mục đích chính của việc phê chuẩn sẽ là giảm khả năng xảy ra lỗi sai mang tính thảm họa. Nhìn chung, việc nhắm tới mục tiêu giảm thiểu nguy cơ sai lỗi mang tính thảm họa thay vì tăng tối đa khả năng tối ưu hóa của mọi chi tiết xem ra khá khôn ngoan. Có hai lý do cho việc này. Đầu tiên, tài nguyên vũ trụ của loài người về mặt thiên văn là rất lớn. Chúng ta có rất nhiều nơi để đi, ngay cả khi quá trình thực hiện của chúng ta gây ra một vài lãng phí hoặc chấp nhận một số hạn chế không cần thiết. Thứ hai, có hy vọng rằng nếu chúng ta chỉ có thể thỏa mãn gần đủ những điều kiện ban đầu cho sự bùng nổ trí tuệ, siêu trí tuệ cuối cùng vẫn có thể tiếp nối và hoàn thành chính xác mục tiêu tối thượng của chúng ta. Điều quan trọng là cần “hạ cánh” đúng chỗ.

Liên quan đến nhận thức luận, có thể một loạt các tiên nghiệm cuối cùng sẽ hội tụ vào các hậu nghiệm tương tự nhau (khi được tính toán bởi một siêu trí tuệ và được tạo điều kiện dựa trên một lượng dữ liệu thực tế). Vì vậy, chúng ta không cần lo lắng về việc có được

một nhận thức luận hoàn toàn chính xác. Chúng ta chỉ cần tránh cung cấp cho AI tiên nghiệm quá cực đoan, khiến AI không có khả năng học được sự thật thiết yếu ngay cả khi lợi ích nhận được là hàng loạt kinh nghiệm và phân tích.⁴⁵

Đối với lý thuyết quyết định, nguy cơ sai lỗi không thể phục hồi xem ra lớn hơn. Chúng ta vẫn có thể hy vọng trực tiếp chỉ định được một lý thuyết quyết định đủ tốt. Một AI siêu trí tuệ có thể chuyển sang áp dụng một lý thuyết quyết định mới bất cứ lúc nào; tuy nhiên, nếu bắt đầu với một lý thuyết quyết định đủ sai, nó có thể không thấy được lý do chuyển đổi. Ngay cả khi một tác tử thấy được những lợi ích của việc có một lý thuyết quyết định khác, việc hiện thực hóa điều này có thể đã quá muộn. Chẳng hạn, một tác tử được thiết kế để từ chối việc bị chiếm đoạt có thể được lợi từ việc răn đe những kẻ sẽ chiếm đoạt. Vì lý do này, các tác tử có thể bị đe dọa có thể chủ động áp dụng tốt một lý thuyết quyết định không thể lợi dụng. Tuy nhiên, khi một tác tử có thể bị chiếm đoạt nhận được lời đe dọa và tin vào nó, thiệt hại sẽ xảy ra.

Khi có nhận thức luận và lý thuyết quyết định phù hợp, chúng ta có thể cố gắng thiết kế hệ thống để thực thi CEV hoặc một số nội dung mục tiêu được chỉ định gián tiếp khác. Một lần nữa, chúng ta lại có thể hy vọng về một sự hội tụ: các cách khác nhau để thực thi một động lực giống như CEV sẽ dẫn đến cùng một kết quả lý tưởng. Khi không có sự hội tụ này, chúng ta vẫn có thể hy vọng rằng sẽ có nhiều kết quả khả thể đủ tốt để được coi là thành công mang tính sống còn.

Chúng ta không cần thiết phải tạo ra một thiết kế được tối ưu hóa cao. Thay vào đó, chúng ta nên tập trung vào việc tạo ra một thiết kế có độ tin cậy cao trong việc duy trì sự tinh tảo để nhận ra những sai

lỗi của mình. Một siêu trí tuệ không hoàn hảo, với các nguyên tắc cơ bản phù hợp, sẽ dần dần tự sửa chữa; và khi đã làm như vậy, nó sẽ phát huy năng lực tối ưu hóa có lợi đối với thế giới như thể nó đã hoàn hảo ngay từ đầu.

Chương 14

Bức tranh chiến lược

Bây giờ là lúc để xem xét thách thức siêu trí tuệ trong bối cảnh rộng hơn. Chúng ta muốn định hướng bản thân trong bối cảnh chiến lược đủ để ít nhất biết được phương hướng khái quát mà mình sẽ nhắm tới. Đây không phải chuyện dễ dàng. Trong chương áp chót này, chúng tôi sẽ giới thiệu một số khái niệm phân tích khái quát giúp suy nghĩ về các vấn đề chính sách khoa học và công nghệ dài hạn. Sau đó chúng tôi sẽ áp dụng những khái niệm này cho vấn đề trí tuệ máy móc.

Việc phân biệt rõ ràng giữa hai quan điểm quy chuẩn khác nhau mà từ đó một chính sách đề xuất được đánh giá có thể soi sáng nhiều vấn đề. Quan điểm ảnh hưởng đến cá nhân đặt ra câu hỏi liệu một sự thay đổi được đề xuất sẽ có phải là “điều chúng ta quan tâm”, hay nói cách khác, (sau khi cân nhắc kỹ và trong kỳ vọng) liệu đó có phải là mối quan tâm của những cá thể đáng cân nhắc về đạo đức, đã tồn tại hay sẽ tồn tại độc lập, bất kể sự thay đổi được đề xuất có xảy ra hay không. Quan điểm phi cá nhân, ngược lại không đưa ra sự cân nhắc đặc biệt nào cho những người đang tồn tại, hoặc sẽ tồn tại độc lập, về việc liệu sự thay đổi được đề xuất có xảy ra hay không. Thay vào đó, quan điểm này cào bằng tất cả, bất kể vị trí của họ trong dòng thời gian. Quan điểm phi cá nhân nhìn thấy giá trị to

lớn trong việc tạo ra những con người mới, nếu cuộc đời của họ là đáng sống: cuộc đời hạnh phúc được tạo ra càng nhiều càng tốt.

Sự khác biệt này tuy hầu như không gợi ý về sự phức tạp đạo đức liên quan đến một cuộc cách mạng trí tuệ máy nhưng có thể hữu ích trong việc phân tích ban đầu. Ở đây, trước tiên chúng ta sẽ xem xét các vấn đề từ quan điểm phi cá nhân. Sau đó, chúng ta sẽ thấy được điều gì thay đổi khi các cân nhắc ảnh hưởng đến con người được tính đến trong quá trình bàn luận.

Chiến lược khoa học kỹ thuật

Trước khi xem xét chi tiết các vấn đề đặc thù của siêu trí tuệ máy, chúng ta cần bàn về một số khái niệm và cân nhắc chiến lược có liên quan khái quát đến việc phát triển khoa học công nghệ.

Nguyên tắc phát triển công nghệ thiên lệch

Giả sử một nhà hoạch định chính sách đề xuất cắt giảm tài trợ cho một lĩnh vực nghiên cứu nhất định vì lo ngại về rủi ro hoặc hậu quả lâu dài của một số công nghệ giả định cuối cùng có thể từ đó đâm chồi. Sau đó, người này có thể mong đợi sự phản đối từ cộng đồng nghiên cứu.

Các nhà khoa học và công chúng ủng hộ họ thường nói rằng thật phù phiếm khi cố gắng kiểm soát sự phát triển của công nghệ bằng cách phong tỏa hoạt động nghiên cứu. Họ lập luận rằng nếu một số công nghệ là khả thi, nó sẽ được phát triển bất kể các nhà hoạch định chính nói gì về những rủi ro mang tính suy đoán trong tương lai. Thật vậy, khi những năng lực mà một hướng phát triển hứa hẹn sẽ tạo ra càng mạnh mẽ, chúng ta càng có thể chắc chắn rằng ai đó, ở đâu đó, sẽ có động lực theo đuổi nó. Việc cắt giảm ngân sách sẽ

không ngăn trở được tiến bộ, và cũng không đón đầu được những mối nguy hiểm đi kèm.

Thú vị thay, sự phản đối này gần như không bao giờ được nêu ra khi một nhà hoạch định chính sách đề xuất tăng tài trợ cho một lĩnh vực nghiên cứu, mặc dù lời lập luận xem ra chống lại cả hai hoạt động tăng vốn và phong tỏa. Chúng ta hiếm khi nghe thấy những tiếng nói phản nộ: “Xin vui lòng đừng tăng tài trợ cho chúng tôi mà hãy thực hiện một số cắt giảm. Các nhà nghiên cứu ở các quốc gia khác chắc chắn sẽ chọn cách thực hiện chậm chạp; và dù thế nào công việc vẫn sẽ hoàn thành. Đừng phung phí của công cho hoạt động nghiên cứu khoa học trong nước!”

Điều gì giải thích cho suy nghĩ nước đôi rõ ràng này? Dĩ nhiên, một lời giải thích hợp lý là các thành viên của cộng đồng nghiên cứu có thiên kiến vị kỷ khiến chúng ta tin rằng nghiên cứu luôn tốt và lôi kéo chúng ta chấp nhận hầu hết mọi lý lẽ ủng hộ yêu cầu cấp thêm vốn. Tuy nhiên, tiêu chuẩn kép này cũng có khả năng được bao biện bằng lợi ích quốc gia. Giả sử việc phát triển công nghệ có hai tác động: mang lại lợi ích nhỏ B cho các nhà phát minh và quốc gia tài trợ cho họ, trong khi gây ra một tác hại H lớn hơn về tổng thể, có thể là ngoại ứng rủi ro (risk externality), đối với mọi người. Từ đó, ngay cả những người có lòng vị tha cũng có thể chọn việc phát triển công nghệ có hại về mặt tổng thể này. Họ có thể lập luận rằng tác hại H sẽ diễn ra bất kể họ làm gì, vì nếu họ kiềm chế, người khác sẽ phát triển công nghệ đó bằng mọi cách. Và trong trường hợp lợi ích tổng thể không thể bị ảnh hưởng, họ cũng có thể nắm lấy lợi ích B cho bản thân và quốc gia của mình. (“Không may rằng sẽ sớm có một thiết bị

phá hủy thế giới. Những cũng thật may, chúng ta đã nhận được tài trợ để xây dựng nó!”)

Dù lời giải thích cho ý kiến phản đối này là gì đi nữa thì nó cũng không thể chứng minh việc không có lý do phi cá nhân để cố gắng định hướng phát triển công nghệ, ngay cả khi chúng ta thừa nhận ý tưởng đầy động lực rằng với những nỗ lực phát triển khoa học công nghệ liên tục, toàn bộ công nghệ liên quan cuối cùng sẽ được phát triển. Hay nói cách khác, ngay cả khi chúng ta thừa nhận điều sau đây:

Giả định hoàn thiện công nghệ:

Nếu các nỗ lực phát triển khoa học công nghệ không hoàn toàn chấm dứt, con người sẽ có được toàn bộ năng lực cơ bản quan trọng có thể có thông qua một số công nghệ khả thể.¹

Có ít nhất hai lý do tại sao giả định hoàn thiện công nghệ không hàm chỉ sự phản đối đã nêu ở trên. Đầu tiên, vẻ thứ nhất của phần mô tả có thể không đúng, vì thực tế là các nỗ lực phát triển công nghệ sẽ không chấm dứt (trước khi đạt được sự hoàn thiện về công nghệ). Sự bảo toàn này đặc biệt thích hợp trong bối cảnh liên quan đến rủi ro diệt vong. Thứ hai, ngay cả khi chúng ta có thể chắc chắn có được tất cả các năng lực cơ bản quan trọng có thể có được thông qua một số công nghệ khả thể thì việc cố gắng gây ảnh hưởng lên hướng nghiên cứu công nghệ vẫn có ý nghĩa nhất định. Điều quan trọng không chỉ là liệu một công nghệ có được phát triển hay không, mà còn là khi nào nó được phát triển, bởi ai và trong bối cảnh nào. Những tình huống ra đời của một công nghệ mới định hình các tác động của nó có thể bị ảnh hưởng bằng cách dừng hoặc gia tăng cấp vốn (và sử dụng các công cụ chính sách khác).

Những điều trên gợi ý về một nguyên tắc sẽ khiến chúng ta chú ý vào tốc độ tương đối giữa sự phát triển của các công nghệ khác nhau.²

Nguyên tắc phát triển công nghệ thiên lệch

Giảm tốc độ phát triển của các công nghệ nguy hiểm và có hại, đặc biệt là những công nghệ làm tăng mức độ rủi ro diệt vong; và đẩy nhanh sự phát triển của các công nghệ có lợi, đặc biệt là những công nghệ làm giảm các rủi ro diệt vong từ tự nhiên hoặc các công nghệ khác.

Do đó, một chính sách có thể được đánh giá dựa trên mức độ lợi thế thiên lệch mà nó mang lại cho các hình thức phát triển công nghệ mong muốn so với các hình thức không mong muốn.³

Thứ tự cán đích ưu tiên

Một số công nghệ có tác động hai chiều lên các rủi ro sống còn, gia tăng mức độ của một số rủi ro trong khi giảm thiểu các rủi ro khác. Siêu trí tuệ là một trong những công nghệ như vậy.

Như chúng ta đã thấy ở các chương trước, sự ra đời của siêu trí tuệ máy sẽ tạo ra một rủi ro diệt vong đáng kể nhưng cũng sẽ giảm thiểu nhiều rủi ro diệt vong khác. Rủi ro từ thiên nhiên, chẳng hạn như tác động của các tiểu hành tinh, siêu núi lửa và đại dịch tự nhiên, hầu như sẽ bị loại bỏ, vì siêu trí tuệ có thể triển khai các biện pháp đối phó với hầu hết các mối nguy đó, hoặc ít nhất là đưa chúng vào danh mục rủi ro không mang tính sống còn (chẳng hạn như thông qua việc thuộc địa hóa không gian).

Những rủi ro diệt vong từ thiên nhiên là tương đối nhỏ khi xét trên thang thời gian phù hợp. Nhưng siêu trí tuệ cũng sẽ loại bỏ hoặc giảm thiểu nhiều rủi ro do con người gây ra. Đặc biệt, nó sẽ giảm

thiếu rủi ro phá hoại ngẫu nhiên, bao gồm rủi ro tai nạn liên quan đến các công nghệ mới. Có năng lực cao hơn con người về tổng thể, một siêu trí tuệ sẽ ít có khả năng mắc lỗi và dễ nhận ra thời điểm cần đến các biện pháp phòng ngừa và thực hiện chúng một cách thành thạo. Một siêu trí tuệ được xây dựng tốt đôi lúc có thể chấp nhận rủi ro, nhưng chỉ khi làm như vậy là khôn ngoan. Hơn nữa, ít nhất là trong các tình huống mà siêu trí tuệ hình thành được một thể đơn nhất, nhiều rủi ro diệt vong do con người chủ ý gây ra xuất phát từ các vấn đề phối hợp toàn cầu sẽ bị loại bỏ. Những rủi ro này bao gồm chiến tranh, chạy đua công nghệ, các hình thức cạnh tranh và tiến hóa không mong muốn, và một hiện tượng được gọi là “bi kịch của tài sản công cộng” (tragedies of the commons)^a.

^a Hiện tượng các tài nguyên chung được sử dụng tự do nên bị khai thác kiệt quệ. (ND)

Vì sự nguy hiểm đáng kể liên quan đến việc con người phát triển sinh học tổng hợp, công nghệ nano phân tử, kỹ thuật khí hậu, các công cụ tăng cường y sinh và thao túng tâm lý-thần kinh, các công cụ kiểm soát xã hội có thể tạo điều kiện cho chế độ toàn trị hoặc chuyên chế, và các công nghệ khác chưa được hình dung, nên việc loại bỏ các rủi ro này là một lợi ích lớn. Vì thế, lập luận có thể đưa ra là chúng ta cần siêu trí tuệ xuất hiện sớm hơn. Tuy nhiên, nếu rủi ro từ thiên nhiên và từ các mối nguy hiểm khác không liên quan đến công nghệ tương lai là nhỏ, thì lập luận này có thể được điều chỉnh: vấn đề nằm ở chỗ chúng ta cần có được siêu trí tuệ trước các công nghệ nguy hiểm khác, như công nghệ nano tiên tiến. Điều đó xảy ra sớm hay muộn có thể không quá quan trọng (theo quan điểm phi cá nhân), miễn là đúng theo thứ tự cán đích.

Cơ sở để ưu tiên siêu trí tuệ đi trước các công nghệ tiềm ẩn nguy hiểm khác, như công nghệ nano, là siêu trí tuệ sẽ làm giảm rủi ro diệt vong từ công nghệ nano chứ không phải điều ngược lại. Do đó, nếu tạo ra siêu trí tuệ trước tiên, chúng ta sẽ chỉ phải đối mặt với những rủi ro diệt vong gắn liền với siêu trí tuệ; trong khi đó, nếu tạo ra công nghệ nano trước, chúng ta sẽ phải đối mặt với những rủi ro từ công nghệ nano và sau đó là cả siêu trí tuệ.⁵ Ngay cả khi rủi ro diệt vong từ siêu trí tuệ là rất lớn, và ngay cả khi siêu trí tuệ mang tính rủi ro cao nhất trong tất cả các công nghệ thì việc đẩy nhanh tốc độ cán đích của nó cũng sẽ là đúng đắn.

Tuy nhiên, lập luận “càng sớm càng tốt” cũng giả định rằng rủi ro của việc tạo ra siêu trí tuệ là như nhau bất kể thời điểm nó được tạo ra. Nếu, thay vào đó, rủi ro giảm dần theo thời gian, có lẽ tốt hơn chúng ta nên trì hoãn cuộc cách mạng trí tuệ máy. Tuy siêu trí tuệ xuất hiện muộn hơn sẽ tạo cơ hội cho các thảm họa sống còn khác xen vào, việc giảm tốc độ phát triển của siêu trí tuệ vẫn có thể là đáng ưu tiên. Điều này sẽ đặc biệt hợp lý nếu các rủi ro diệt vong liên quan đến siêu trí tuệ lớn hơn nhiều so với rủi ro liên quan đến các công nghệ đột phá khác.

Có một số lý do khá phù hợp để tin rằng rủi ro của sự bùng nổ trí tuệ sẽ giảm đáng kể khi nhiều thập kỷ trôi qua. Một trong số đó là việc xuất hiện muộn sẽ mang lại nhiều thời gian để phát triển các giải pháp cho vấn đề kiểm soát. Vấn đề kiểm soát này chỉ mới được công nhận gần đây, và hầu hết các ý tưởng tốt nhất hiện nay về cách tiếp cận nó cũng chỉ vừa được phát hiện trong vòng một thập kỷ qua (và trong khi cuốn sách này được viết, với một số trường hợp). Có thể công nghệ hiện đại sẽ tiến bộ rất nhiều trong vài thập kỷ tới; và nếu

vấn đề trở nên khó khăn, một phần đáng kể của quá trình này có thể sẽ kéo dài suốt thế kỷ tới. Càng mất nhiều thời gian để siêu trí tuệ xuất hiện, càng có nhiều tiến bộ như vậy được hiện thực hóa. Điều này hết sức ủng hộ việc trì hoãn sự xuất hiện của siêu trí tuệ và phản đối việc để siêu trí tuệ cán đích sớm hơn.

Một lý do khác cho việc siêu trí tuệ đến muộn có thể an toàn hơn là điều này sẽ mang lại nhiều thời gian hơn cho nhiều xu hướng có lợi của nền văn minh nhân loại phát huy tác dụng của mình. Mức độ tin tưởng của chúng ta với ý kiến cân nhắc này sẽ phụ thuộc vào sự lạc quan của chúng ta dành cho các xu hướng đó.

Một người lạc quan chắc chắn có thể chỉ ra một số chỉ số đáng khích lệ cũng như một số khả năng đầy hy vọng. Mọi người có thể học cách sống hòa đồng hơn, giảm bạo lực, chiến tranh và sự tàn ác. Bên cạnh đó, sự phối hợp toàn cầu và phạm vi hội nhập chính trị có thể sẽ tăng lên, giúp dễ dàng thoát khỏi các cuộc đua công nghệ không mong muốn (điều này sẽ được bàn thêm phía dưới) đồng thời thực hiện một thỏa thuận mà theo đó, những lợi ích kỳ vọng từ sự bùng nổ trí tuệ sẽ được chia sẻ rộng rãi. Dường như chúng ta đã có những xu hướng lịch sử diễn ra trong thời gian dài theo hướng đi này.⁶

Hơn nữa, một người lạc quan cũng có thể hy vọng rằng “mức độ đúng đắn” của con người sẽ tăng lên trong suốt thế kỷ này, vì những định kiến sẽ giảm đi (sau khi cân nhắc kỹ lưỡng), hiểu biết sẽ được tích lũy và mọi người sẽ quen với việc nghĩ về những khả năng mang tính trừu tượng trong tương lai cũng như những rủi ro toàn cầu. Nếu may mắn, chúng ta có thể thấy một sự cải thiện chung của các tiêu chuẩn nhận thức trong cả nhận thức cá nhân và tập thể. Một

lần nữa, ta lại có những xu hướng tiến triển theo hướng đi này. Tiến bộ khoa học đồng nghĩa với việc hiểu biết sẽ gia tăng. Tăng trưởng kinh tế có thể cung cấp đầy đủ dinh dưỡng cho một phần lớn dân số thế giới (đặc biệt là trong những năm đầu đời quan trọng đối với sự phát triển của não bộ) và tạo cơ hội tiếp cận với nền giáo dục có chất lượng. Những tiến bộ trong công nghệ thông tin sẽ giúp bạn dễ dàng tìm kiếm, kết hợp, đánh giá và truyền đạt dữ liệu và ý tưởng. Hơn nữa, vào cuối thế kỷ, nhân loại sẽ tiếp tục mắc sai lầm thêm 100 năm, và từ đó có thể học thêm được nhiều điều.

Theo những gì đã nói, nhiều sự phát triển tiềm năng sẽ có tính hai chiều, làm gia tăng một vài rủi ro diệt vong và giảm thiểu những rủi ro khác. Chẳng hạn, những tiến bộ trong giám sát, khai thác dữ liệu, phát hiện nói dối, sinh trắc học và các phương tiện tâm lý hoặc hóa- thần kinh để điều khiển niềm tin và ham muốn có thể giảm thiểu một vài rủi ro diệt vong nhờ tạo điều kiện phối hợp quốc tế hoặc áp chế những thành phần khủng bố và phạm pháp. Tuy nhiên, những tiến bộ này cũng có thể làm gia tăng một số rủi ro diệt vong với việc khuếch đại các động lực xã hội không mong muốn hoặc cho phép hình thành các chế độ toàn trị ổn định vĩnh viễn.

Một ranh giới quan trọng là sự tăng cường nhận thức sinh học, chẳng hạn như thông qua chọn lọc di truyền. Khi thảo luận điều này trong Chương 2 và 3, chúng ta đã kết luận rằng các hình thức siêu trí tuệ hoàn thiện nhất nhiều khả năng sẽ phát sinh dưới dạng trí tuệ máy. Nhận định nhất quán với việc tăng cường nhận thức đóng vai trò quan trọng trong việc mở đường và tạo ra siêu trí tuệ máy. Tăng cường nhận thức xem ra rõ ràng có thể giảm thiểu rủi ro: những người nghiên cứu vấn đề kiểm soát càng thông minh thì càng có

hiều khả năng tìm ra giải pháp. Tuy nhiên, việc tăng cường nhận thức cũng có thể thúc đẩy sự phát triển của trí tuệ máy, do đó làm giảm quỹ thời gian sẵn có dành cho việc giải quyết vấn đề. Tăng cường nhận thức cũng sẽ có nhiều hậu quả liên quan khác, và những vấn đề này phải được xem xét kỹ hơn. (Hầu hết các nhận xét sau đây về “tăng cường nhận thức” áp dụng như nhau đối với các phương tiện phi sinh học giúp cải thiện hiệu quả nhận thức cá nhân hoặc tập thể.)

Tốc độ thay đổi và tăng cường nhận thức

Sự gia tăng về mức trung bình hay mức trần của trí tuệ con người đều có khả năng thúc đẩy tiến bộ công nghệ về mọi mặt, bao gồm tiến bộ về các dạng trí tuệ máy khác nhau, tiến bộ trong vấn đề kiểm soát và tiến bộ trên các mục tiêu kinh tế và kỹ thuật khác. Vậy hiệu ứng cuối cùng của sự gia tốc này là gì?

Hãy xem xét trường hợp một “yếu tố gia tốc phổ quát”, một sự can thiệp tưởng tượng giúp tăng tốc mọi thứ theo nghĩa đen. Hành động của một yếu tố gia tốc phổ quát như vậy sẽ chỉ tương ứng với việc thay đổi về thời gian chứ không tạo ra sự thay đổi về chất trong những kết quả quan sát được.⁷

Nếu muốn hiện thực hóa ý tưởng tăng cường nhận thức tổng thể có thể tăng tốc mọi thứ, chúng ta rõ ràng sẽ cần một số khái niệm khác chứ không phải chỉ là sự tăng tốc phổ quát. Một cách tiếp cận hứa hẹn hơn là tập trung vào việc tăng cường nhận thức có thể làm tăng tốc độ thay đổi trong một loại quy trình như thế nào so với trong một số loại quy trình khác. Một sự tăng tốc thiên lệch như vậy có thể ảnh hưởng đến động lực của hệ thống. Vì vậy, hãy xem xét khái niệm sau:

Yếu tố gia tốc phát triển cấu trúc vĩ mô: Một đòn bẩy giúp gia tăng tốc độ phát triển các đặc điểm cấu trúc vĩ mô của hoàn cảnh nhân loại mà không thay đổi tốc độ tiến triển của các vấn đề con người ở cấp độ vi mô.

Hãy tưởng tượng việc “kéo” đòn bẩy này theo hướng giảm tốc. Một “má phanh” sẽ được đặt vào bánh xe vĩ đại của lịch sử thế giới; tia lửa sẽ bắn ra kèm với tiếng cọt xát của kim loại. Sau khi bánh xe này ổn định hơn, kết quả sẽ là một thế giới mà trong đó, quá trình đổi mới công nghệ xảy ra chậm hơn và những thay đổi cơ bản hoặc toàn cầu trong cấu trúc chính trị và văn hóa xảy ra ít thường xuyên và ít đột ngột hơn. Rất nhiều thế hệ sẽ sinh ra và lụi tàn trước khi một thời đại nhường chỗ cho thời đại khác. Trong suốt cuộc đời, con người sẽ nhìn thấy rất ít thay đổi về cấu trúc cơ bản của hoàn cảnh nhân loại.

Với sự tồn tại của đa phần nhân loại, sự phát triển cấu trúc vĩ mô trước đây đã xảy ra chậm hơn bây giờ. 50 ngàn năm trước, cả một thiên niên kỷ có thể trôi qua mà không có phát minh công nghệ quan trọng, không có sự gia tăng đáng kể về kiến thức và hiểu biết của con người, và không có bất kỳ thay đổi chính trị nào có ý nghĩa toàn cầu. Tuy nhiên, ở cấp độ vi mô, chiếc “kính vạn hoa” về các vấn đề của con người đã xoay vần đáng kể trong sinh, lão, bệnh, tử cùng nhiều sự kiện quan trọng khác của người và đất. Ở thế Pleistocene, người ta có thể trải qua nhiều sự kiện trong ngày hơn so với chúng ta hiện nay.

Nếu gặp một đòn bẩy ma thuật cho phép thay đổi tốc độ phát triển cấu trúc vĩ mô, bạn nên làm gì? Bạn phải tăng tốc, giảm tốc hay để mọi thứ diễn ra như hiện tại?

Theo quan điểm phi cá nhân, câu hỏi này cần chúng ta xem xét các tác động lên mức độ rủi ro diệt vong. Hãy phân biệt hai loại rủi ro: “rủi ro trạng thái” và “rủi ro đơn đoạn”. Rủi ro trạng thái là rủi ro liên quan đến việc có một trạng thái nhất định và số lượng rủi ro trạng thái mà hệ thống phải đối mặt sẽ tỷ lệ thuận với thời gian mà hệ thống duy trì trạng thái đó. Rủi ro từ tự nhiên thường là rủi ro trạng thái: chúng sơ hở càng lâu thì càng dễ bị tấn công bởi, chẳng hạn, một tiểu hành tinh, một siêu núi lửa, một vụ nổ tia gamma, một đại dịch phát sinh tự nhiên hoặc một thảm họa khác từ vũ trụ. Một số rủi ro nhân tạo cũng là rủi ro trạng thái. Ở cấp độ cá nhân, một người lính ló đầu lên trên lan can càng lâu, khả năng anh ta bị kẻ thù bắn tã càng lớn. Một số rủi ro trạng thái do con người tạo ra cũng có thể lên đến mức sống còn: chúng ta càng sống lâu trong một hệ thống vô chính phủ toàn cầu, khả năng xảy ra một cuộc chiến nhiệt hạch sống còn hoặc một cuộc chiến tranh lớn với các loại vũ khí huỷ diệt hàng loạt khác sẽ huỷ diệt toàn bộ nền văn minh nhân loại.

Rủi ro đơn đoạn, ngược lại, là rủi ro riêng biệt liên quan đến một quá trình chuyển đổi cần thiết hoặc mong muốn nào đó. Khi quá trình chuyển đổi hoàn thành, rủi ro sẽ biến mất. Số lượng rủi ro đơn đoạn liên quan đến quá trình chuyển đổi thường không chỉ liên quan đến thời gian chuyển đổi. Người ta không thể giảm một nửa rủi ro của việc vượt qua một bãi mìn bằng cách chạy nhanh gấp đôi. Trong điều kiện cất cánh nhanh, việc tạo ra siêu trí tuệ có thể là rủi ro đơn đoạn: sẽ có một mức độ rủi ro nhất định liên quan đến việc cất cánh phụ thuộc vào việc chuẩn bị đã được thực hiện, nhưng có thể không phụ thuộc nhiều vào việc cất cánh mất 20 giây hay 20 giờ.

Vì thế, chúng ta có thể nói về một yếu tố gia tốc phát triển cấu trúc vĩ mô giả thuyết như sau:

- Khi lo ngại về rủi ro trạng thái có tính sống còn, chúng ta nên ưu tiên tăng tốc, với điều kiện chúng ta nghĩ rằng mình có triển vọng thực tế để cầm cự đến thời kỳ hậu chuyển đổi, khi mọi rủi ro diệt vong sẽ giảm đi rất nhiều.

- Nếu biết rằng một thảm họa sống còn chắc chắn sẽ xảy ra, chúng ta phải giảm tốc độ phát triển cấu trúc vĩ mô (hoặc thậm chí đảo ngược tiến trình) để tạo cơ hội tồn tại cho nhiều thế hệ trước khi “hạ màn”. Nhưng trên thực tế, sẽ là quá bi quan nếu quả quyết rằng nhân loại đang đi đến hồi kết.

- Hiện tại, mức độ rủi ro trạng thái mang tính sống còn đang khá thấp. Nếu tưởng tượng rằng các điều kiện vĩ mô về công nghệ khiến nhân loại bị đóng băng trong trạng thái hiện tại, có vẻ như rất khó xảy ra thảm họa tồn tại trong khoảng thời gian, giả sử, một thập kỷ. Vì vậy, thời gian trễ một thập kỷ, nếu xảy ra ở giai đoạn phát triển hiện tại hoặc vào một thời điểm khác khi rủi ro trạng thái ở mức thấp, sẽ chỉ mang đến một rủi ro diệt vong không đáng kể. Trong khi đó, sự trì hoãn một thập kỷ đối với những sự phát triển công nghệ kế tiếp có thể cũng có tác động có lợi đáng kể lên những rủi ro đơn đoạn mang tính sống còn sau này, chẳng hạn như kéo dài thời gian chuẩn bị.

Kết quả: tầm quan trọng của tốc độ phát triển cấu trúc vĩ mô chủ yếu nằm ở ảnh hưởng của nó đến mức độ chuẩn bị của con người vào thời điểm phải đương đầu với những rủi ro đơn đoạn quan trọng.⁸

Vì vậy, câu hỏi chúng ta phải đặt ra là việc tăng cường nhận thức (và đồng thời tăng tốc quá trình phát triển cấu trúc vĩ mô) sẽ ảnh

hưởng như thế nào đến mức độ chuẩn bị mong muốn tại thời điểm quan trọng. Chúng ta có nên ưu tiên một thời gian chuẩn bị ngắn hơn với trí tuệ cao hơn? Với trí tuệ cao hơn, thời gian chuẩn bị có thể được sử dụng hiệu quả hơn, và bước quan trọng cuối cùng sẽ được một cộng đồng nhân loại thông minh hơn thực hiện. Hoặc liệu chúng ta có nên ưu tiên một mức độ gần với trí tuệ hiện tại hơn nếu điều đó cho chúng ta thêm thời gian chuẩn bị?

Lựa chọn nào tốt hơn phụ thuộc vào bản chất của thử thách mà chúng ta chuẩn bị đương đầu. Nếu thử thách là giải quyết một vấn đề mà với đó, học hỏi kinh nghiệm là then chốt, thì độ dài thời gian của giai đoạn chuẩn bị có thể là yếu tố quyết định vì chúng ta cần thời gian để tích lũy kinh nghiệm cần thiết. Thách thức này cụ thể sẽ là gì? Giả sử, đó là một công nghệ vũ khí mới mà chúng ta dự đoán sẽ được phát triển vào một thời điểm nào đó trong tương lai, và điều này chắc chắn sẽ dẫn đến việc chiến tranh sau đó sẽ có xác suất 1/10 gây ra một thảm họa sống còn. Nếu đó là bản chất của thách thức, chúng ta có thể mong muốn tốc độ phát triển cấu trúc vĩ mô chậm lại để kịp thời cùng nhau hành động trước khi công nghệ vũ khí mới này ra đời. Ai đó có thể hy vọng rằng trong khoảng thời gian gia tăng được bảo đảm bằng sự giảm tốc, nhân loại có thể tìm ra cách tránh chiến tranh, và các mối quan hệ quốc tế khi đó có thể sẽ giống như các nước EU, vốn đã đấu đá kịch liệt hàng thế kỷ, nay cùng chung sống hòa bình. Hòa bình có thể có được nhờ hàng loạt các quá trình khai hóa văn minh, hoặc một cú sốc đến từ các thảm họa gần như sống còn (chẳng hạn như các vụ nổ hạt nhân nhỏ, và hoạt động đàm phán mà chúng có thể mang lại để cuối cùng tạo ra các thể chế toàn cầu cần thiết nhằm hòa giải chiến tranh giữa các quốc

gia). Nếu hình thức học hoặc điều chỉnh này không được tăng tốc đáng kể nhờ cải thiện trí tuệ, việc tăng cường nhận thức sẽ là điều không mong muốn và chỉ giúp “châm ngòi” rủi ro nhanh hơn mà thôi.

Tuy nhiên, một sự bùng nổ trí tuệ tiềm tàng có thể đem lại một kiểu thách thức khác. Vấn đề kiểm soát đòi hỏi tầm nhìn xa, lý luận và hiểu biết về lý thuyết. Ích lợi của việc gia tăng kinh nghiệm lịch sử là không rõ ràng. Chúng ta không thể trực tiếp tích lũy kinh nghiệm từ sự bùng nổ trí tuệ (cho đến khi quá muộn), và nhiều yếu tố kết hợp lại khiến vấn đề kiểm soát là độc nhất và không có tiền lệ. Vì những lý do này, lượng thời gian trôi qua trước khi xảy ra sự bùng nổ trí tuệ có thể sẽ không mấy quan trọng. Có lẽ, điều quan trọng, thay vào đó, là (a) mức độ tiến bộ trí tuệ trên vấn đề kiểm soát đạt được tại thời điểm bùng nổ; và (b) lượng kỹ năng và trí tuệ sẵn có tại thời điểm đó để thực hiện các giải pháp khả dụng tốt nhất (và để ứng biến với những gì còn thiếu).⁹ Yếu tố (b) hiển nhiên sẽ chịu tác động của việc tăng cường nhận thức, nhưng tăng cường nhận thức sẽ ảnh hưởng đến yếu tố (a) như thế nào lại là một vấn đề phức tạp hơn.

Như đã nói, tăng cường nhận thức sẽ là yếu tố gia tốc phát triển cấu trúc vĩ mô tổng thể. Điều này sẽ đẩy nhanh quá trình xuất hiện của sự bùng nổ trí tuệ, do đó làm giảm thời gian cho việc chuẩn bị và đạt được tiến bộ trong vấn đề kiểm soát. Thông thường đây sẽ là một điều xấu. Tuy nhiên, nếu lý do duy nhất của việc có ít thời gian dành cho tiến bộ trí tuệ là tiến bộ trí tuệ được tăng tốc, chúng ta cần ngăn chặn mọi sự suy giảm mức độ tiến bộ trí tuệ sẽ có khi sự bùng nổ trí tuệ xảy ra.

Tại thời điểm này, tăng cường nhận thức xem ra không tác động lên yếu tố (a): trí tuệ vẫn sẽ tiến bộ với mức độ tương đương, bao

gồm cả tiến bộ trong việc giải quyết vấn đề kiểm soát, và chỉ được rút gọn lại trong một khoảng thời gian ngắn hơn. Tuy nhiên, trong thực tế, tăng cường nhận thức cũng rất có thể có ảnh hưởng tích cực đến (a).

Một lý do để tăng cường nhận thức có thể gây ra nhiều tiến bộ hơn trong việc giải quyết vấn đề kiểm soát vào thời điểm sự bùng nổ trí tuệ xảy ra là sự tiến bộ trong vấn đề kiểm soát có thể đặc biệt phụ thuộc các mức độ hiệu năng trí tuệ cực đại, thậm chí còn nhiều hơn so với việc cần làm để tạo ra trí tuệ máy. Vai trò của phương pháp thử và sai cùng quá trình tích lũy kết quả thực nghiệm xem ra là khá hạn chế so với vấn đề kiểm soát, trong khi phương pháp học thực nghiệm (experimental learning) có thể sẽ đóng một vai trò lớn trong việc phát triển trí tuệ nhân tạo hoặc giả lập hoàn chỉnh não bộ. Vì thế giữa các tác vụ, thời gian sẽ bù trừ cho trí tuệ ở mức độ khác nhau, và sự khác nhau này nên khiến tăng cường nhận thức thúc đẩy tiến bộ trong việc giải quyết vấn đề kiểm soát nhiều hơn so với vấn đề chế tạo trí tuệ máy.

Một lý do khác của việc tăng cường nhận thức nên thúc đẩy một cách thiên lệch sự tiến bộ trong vấn đề kiểm soát nằm ở chỗ, nhu cầu đối với sự tiến bộ như vậy rất có thể sẽ được coi trọng bởi các cộng đồng xã hội và cá nhân có khả năng nhận thức cao hơn. Cần có tầm nhìn xa và lý luận để nhận ra tại sao vấn đề kiểm soát lại quan trọng và đặt nó làm ưu tiên.¹⁰ Một sự khôn ngoan xuất chúng có thể cũng cần đến để tìm ra những cách thức hứa hẹn nhằm tiếp cận một vấn đề lạ lẫm như vậy.

Từ những vấn đề này, chúng ta có thể tạm kết luận rằng việc tăng cường nhận thức là đáng mong đợi, ít nhất là khi chúng ta đang

chú tâm vào những rủi ro diệt vong của một sự bùng nổ trí tuệ. Ý tưởng tương tự cũng áp dụng cho các rủi ro diệt vong khác phát sinh từ những thách thức đòi hỏi phải có tầm nhìn xa và lý luận trừu tượng đáng tin cậy (trái ngược với, chẳng hạn, sự thích ứng tăng dần trước những thay đổi đã được trải nghiệm trong môi trường hoặc quá trình đúc kết văn hóa và xây dựng thể chế kéo dài nhiều thế hệ).

Khớp nối công nghệ

Giả sử ai đó nghĩ rằng việc giải quyết vấn đề kiểm soát với trí tuệ nhân tạo là rất khó, giải quyết nó với giả lập hoàn chỉnh não bộ dễ dàng hơn nhiều, và tốt hơn là nên tạo ra trí tuệ máy thông qua con đường giả lập hoàn chỉnh não bộ. Chúng ta sẽ quay lại sau với câu hỏi liệu giả lập hoàn chỉnh não bộ có an toàn hơn trí tuệ nhân tạo hay không, nhưng bây giờ chúng tôi muốn đưa ra quan điểm rằng ngay cả khi chúng ta chấp nhận tiền đề này, thì điều đó cũng không kéo theo việc chúng ta phải thúc đẩy công nghệ giả lập hoàn chỉnh não bộ. Một lý do đã được thảo luận trước đây là sự xuất hiện muộn hơn của siêu trí tuệ có thể sẽ tốt hơn. Điều này mang đến nhiều thời gian hơn để giải quyết vấn đề kiểm soát và khiến các xu hướng nền tảng thuận lợi khác đạt đến đỉnh cao, và do đó, nếu người ta tin rằng dù thế nào đi nữa, giả lập hoàn chỉnh não bộ vẫn sẽ đi trước AI, thì việc thúc đẩy hơn nữa quá trình tạo ra giả lập hoàn chỉnh não bộ sẽ phản tác dụng.

Nhưng ngay cả khi giả lập hoàn chỉnh não bộ cần xuất hiện càng sớm càng tốt, chúng ta vẫn không nên theo đuổi tiến trình này. Vì có thể tiến trình giả lập hoàn chỉnh não bộ sẽ không mang lại bất kỳ bản giả lập nào, mà thay vào đó, nó có thể mang đến trí tuệ nhân tạo hình

thái thần kinh (các dạng AI sao chép một số đặc điểm tổ chức của vỏ não nhưng không tái tạo chức năng thần kinh với độ trung thực đủ để tạo thành một bản giả lập thích hợp). Nếu (và có lý do để tin rằng) AI có cấu trúc thần kinh như vậy tệ hơn loại AI được tạo ra, và nếu bằng cách thúc đẩy giả lập hoàn chỉnh não bộ, chúng ta sẽ khiến AI hình thái thần kinh cán đích trước, thì việc chúng ta theo đuổi kết quả tốt nhất (giả lập hoàn chỉnh não bộ) sẽ dẫn đến kết quả tồi tệ nhất (AI hình thái thần kinh); còn nếu theo đuổi kết quả tốt thứ hai (AI tổng hợp), chúng ta thực sự có thể đạt được nó.

Trên đây là đoạn mô tả một ví dụ (giả thuyết) về những gì có thể được gọi là “khớp nối công nghệ”.¹¹ Đây là điều kiện mà trong đó hai công nghệ có mối quan hệ khả đoán về thời điểm, sao cho việc phát triển một trong những công nghệ sẽ rất dễ dẫn đến sự phát triển của công nghệ kia, giống như một công nghệ tiên quyết, hoặc một ứng dụng hay bước đi kế tiếp rõ ràng và bất khả kháng. Các khớp nối công nghệ phải được tính đến khi chúng ta sử dụng nguyên tắc phát triển công nghệ thiên lệch: thúc đẩy sự phát triển của công nghệ Y mong muốn sẽ không có lợi nếu cách duy nhất để có được Y là phát triển một công nghệ tiền thân X cực kỳ không mong muốn, hoặc nếu có Y, một công nghệ liên quan cực kỳ không mong muốn Z sẽ được tạo ra. Người xưa thường nói, “lấy vợ xem tông, lấy chồng xem giống”.

Đối với trường hợp giả lập hoàn chỉnh não bộ, mức độ khớp nối công nghệ là điều gây tranh cãi. Như đã lưu ý trong Chương 2, mặc dù giả lập hoàn chỉnh não bộ đòi hỏi tiến bộ lớn trong các công nghệ mở đường, nhưng nó có thể không yêu cầu bất kỳ hiểu biết lý thuyết mới nào đáng kể. Cụ thể, chúng ta không cần hiểu cách thức hoạt

động của nhận thức con người, mà chỉ cần biết cách xây dựng các mô hình tính toán của các phần nhỏ trong não, chẳng hạn như các loại neuron khác nhau. Tuy nhiên, trong quá trình phát triển khả năng giả lập não bộ con người, rất nhiều dữ liệu về thần kinh sẽ được thu thập, và các mô hình chức năng của mạng lưới vỏ não chắc chắn sẽ được cải thiện rất nhiều. Sự tiến bộ như vậy xem ra có nhiều khả năng kích hoạt AI hình thái thần kinh trước khi giả lập hoàn chỉnh não bộ thành công.¹² Trong lịch sử, có khá nhiều ví dụ về kỹ thuật AI lượm lặt từ khoa học thần kinh hoặc sinh học (chẳng hạn: neuron McCulloch-Pitts, các perceptron, các neuron và mạng lưới thần kinh nhân tạo khác, lấy cảm hứng từ các công trình giải phẫu thần kinh; học tăng cường lấy cảm hứng từ tâm lý học hành vi; thuật toán di truyền lấy cảm hứng từ thuyết tiến hóa; kiến trúc gộp và hệ thống phân cấp tri giác lấy cảm hứng từ các lý thuyết khoa học nhận thức về lập kế hoạch vận động và tri giác; hệ miễn dịch nhân tạo, lấy cảm hứng từ miễn dịch học lý thuyết, trí tuệ bầy đàn, lấy cảm hứng từ đặc điểm sinh thái học của các đàn kiến hoặc các hệ thống tự tổ chức khác; và phương pháp điều khiển phản ứng và dựa hành vi trong robot học, lấy cảm hứng từ nghiên cứu sự vận động của các loài). Có lẽ đáng kể hơn, có rất nhiều câu hỏi quan trọng liên quan đến AI có thể được trả lời thông qua việc nghiên cứu sâu hơn về bộ não. (Ví dụ: Bộ não lưu trữ các phép biểu diễn có cấu trúc trong bộ nhớ làm việc và bộ nhớ dài hạn như thế nào? Vấn đề liên kết được giải quyết ra sao? Mã thần kinh là gì? Các khái niệm được biểu diễn như thế nào? Bộ máy xử lý trong vỏ não có đơn vị tiêu chuẩn, chẳng hạn như cột vỏ não hay không, và nếu có thì nó được kết nối ra sao và chức

năng của nó phụ thuộc vào hệ thống kết nối như thế nào? Các cột như vậy được liên kết ra sao? Chúng học như thế nào?)

Chúng ta sẽ sớm có nhiều điều để nói về mức độ nguy hiểm tương quan của giả lập hoàn chỉnh não bộ, AI hình thái thần kinh và AI tổng hợp, nhưng chúng ta đã có thể nhận ra một sự khớp nối công nghệ quan trọng khác: đó là giữa giả lập hoàn chỉnh não bộ và AI. Ngay cả khi việc thúc đẩy giả lập hoàn chỉnh não bộ thực sự cho ra đời một bản giả lập (chứ không phải AI hình thái thần kinh), và ngay cả khi giả lập hoàn chỉnh não bộ có thể được xử lý một cách an toàn thì chúng ta vẫn còn một nguy cơ sâu xa hơn liên quan đến quá trình chuyển đổi thứ hai, một quá trình chuyển đổi từ giả lập hoàn chỉnh não bộ sang AI, một dạng trí tuệ máy mạnh mẽ hơn.

Có nhiều khớp nối công nghệ khác có thể được xem xét trong một quá trình phân tích toàn diện hơn, chẳng hạn, việc thúc đẩy giả lập hoàn chỉnh não bộ sẽ tăng cường sự tiến bộ của khoa học thần kinh nói chung.¹³ Điều đó có thể tạo ra các hiệu ứng khác nhau, chẳng hạn như tiến bộ nhanh hơn nhằm tới phát hiện nói dối, kỹ thuật điều khiển tâm lý thần kinh, tăng cường nhận thức và tiến bộ y tế. Tương tự, việc thúc đẩy tăng cường nhận thức có thể (tùy theo con đường cụ thể) tạo ra các hiệu ứng lan tỏa, chẳng hạn như việc đẩy nhanh tốc độ phát triển phương pháp chọn lọc di truyền hay việc kỹ thuật di truyền không chỉ tăng cường nhận thức mà còn sửa đổi các tính trạng khác.

Phán đoán

Chúng ta sẽ đối mặt với sự phức tạp chiến lược ở một mức độ khác nếu tính đến việc không có người kiểm soát thế giới có bản chất rộng lượng, lý trí và thống nhất, chỉ đơn giản thực thi những gì

đã được biết đến như là lựa chọn tốt nhất. Bất kỳ luận điểm trừu tượng nào về “những gì nên làm” đều phải được thể hiện dưới dạng thông điệp cụ thể, được đưa vào một môi trường thực tế hoa mỹ và mang màu sắc chính trị. Ở đó, nó sẽ bị tảng lờ, hiểu sai, bóp méo hoặc chiếm đoạt cho các mục đích xung đột khác nhau; nó sẽ nảy xung quanh như một quả bóng, gây ra các hành động và phản ứng, báo hiệu một loạt hệ quả không chịu tác động của mối liên hệ trực tiếp với những chủ đích của người gửi thông điệp ban đầu.

Một người tinh tế có thể cố gắng phán đoán các loại hiệu ứng như vậy. Chẳng hạn, hãy xem xét mẫu lập luận sau đây phục vụ việc tiến hành nghiên cứu để phát triển công nghệ nguy hiểm X. (Một lập luận phù hợp với mẫu này có thể được tìm thấy trong các bài viết của Eric Drexler. Trong trường hợp của Drexler, X = công nghệ nano phân tử.¹⁴)

1. Những rủi ro của X là rất lớn.
2. Việc giảm thiểu những rủi ro này đòi hỏi một giai đoạn chuẩn bị nghiêm túc.
3. Việc chuẩn bị nghiêm túc sẽ chỉ bắt đầu khi viễn cảnh của X được xem xét một cách nghiêm túc bởi các tầng lớp chính trong xã hội.
4. Các tầng lớp chính trong xã hội sẽ chỉ xem xét nghiêm túc viễn cảnh của X khi một nỗ lực nghiên cứu lớn nhằm phát triển X đang được tiến hành.
5. Một nỗ lực nghiên cứu nghiêm túc được bắt đầu càng sớm thì càng mất nhiều thời gian hơn để tạo ra X (vì nó khởi đầu từ các công nghệ mở đường thô sơ hơn).

6. Vì thế, một nỗ lực nghiên cứu nghiêm túc được bắt đầu càng sớm thì thời gian chuẩn bị nghiêm túc sẽ càng kéo dài và việc giảm thiểu rủi ro càng lớn.

7. Vì thế, một nỗ lực nghiên cứu nghiêm túc với X nên được lập tức bắt đầu.

Theo lối suy nghĩ này, thứ lúc đầu giống như lý do cho việc giảm tốc độ hoặc ngăn chặn việc X gây ra các rủi ro lớn cuối cùng lại trở thành lý do cho kết luận ngược lại.

Một lối lập luận có liên quan cho rằng chúng ta nên đón nhận những thảm họa quy mô vừa và nhỏ, vì chúng khiến chúng ta nhận thức được các thiếu sót của mình và thúc đẩy chúng ta thực hiện các biện pháp phòng ngừa làm giảm khả năng xảy ra thảm họa. Một thảm họa quy mô nhỏ hoặc trung bình hoạt động giống như tiêm chủng, thử thách nền văn minh với một hình thức đe dọa ở mức có thể sống sót và kích thích phản ứng miễn dịch để sẵn sàng đối phó với hàng loạt mối đe dọa sống còn.¹⁵

Những lập luận kiểu “phản ứng khi bị kích động” này ủng hộ việc tạo ra một tác động xấu với hy vọng nó sẽ kích thích phản ứng đại chúng. Chúng tôi nhắc đến những lập luận này không vì mục đích thực chứng, mà như một cách để giới thiệu ý tưởng về (điều mà chúng tôi gọi là) “những lập luận phán đoán” (second-guessing arguments). Những lập luận như vậy cho rằng bằng cách coi người khác là không hợp lý và công kích vào những thiên kiến cùng quan niệm sai lầm của họ, ta có thể kích thích một phản ứng hiệu quả hơn so với việc đưa ra tình huống một cách thẳng thắn và trung thực cho lý trí của họ.

Xem ra chiến lược mà những lập luận phán đoán đưa ra là rất khó áp dụng để đạt được những mục tiêu toàn cầu dài hạn. Làm sao chúng ta có thể đoán được nội dung cuối cùng của một thông điệp khi nó đã bị xáo xáo trong mớ hỗn loạn của những cuộc tranh luận công khai? Để làm được điều này, xem ra chúng ta phải dự đoán các “hiệu ứng tu từ” đối với vô số thành phần với các tính chất riêng biệt và mức độ ảnh hưởng dao động trong thời gian dài mà qua đó, hệ thống có thể bị nhiễu loạn bởi các sự kiện không lường trước được từ bên ngoài, trong khi cấu trúc liên kết của nó cũng đang trải qua một quá trình tổ chức nội sinh liên tục: đây chắc chắn là một nhiệm vụ bất khả thi!¹⁶ Tuy nhiên, có thể không cần thiết phải đưa ra dự đoán chi tiết về toàn bộ quỹ đạo trong tương lai của hệ thống để xác định một hoạt động can thiệp đáng mong đợi trong việc gia tăng xác suất của một kết quả dài hạn nào đó. Chẳng hạn, người ta có thể xem xét chi tiết các tác động tương đối gần và có thể dự đoán rồi chọn ra một hành động phù hợp, trong khi chỉ lập được mô hình về hành vi khó phán đoán của hệ thống một cách ngẫu nhiên.

Tuy nhiên, việc giảm nhẹ hoặc kiềm chế các động thái phán đoán có thể gặp phải một vấn đề đạo đức. Việc cố gắng lừa gạt lẫn nhau giống như một trò chơi tổng bằng 0 (hoặc tổng âm)^a khi xét đến thời gian và năng lượng sẽ mất đi khi thực hiện, cũng như khả năng của việc nó sẽ khiến mọi người khó mà hiểu được suy nghĩ thật của người khác, cũng như khó được tin tưởng khi bày tỏ quan điểm cá nhân.¹⁷

^a *Trò chơi tổng bằng 0 (zero-sum game): tình huống người thắng được lợi bằng với mất mát của người thua. Trò chơi tổng âm (negative-sum game): cả hai bên đều chịu mất mát. (ND)*

Việc triển khai tối đa các phương thức truyền thông chiến lược sẽ “giết chết” sự ngay thẳng và khiến cho sự thật rơi vào đêm đen với những cú đâm chọc đến từ những bóng ma chính trị.

Lối đi và yếu tố dẫn đường

Chúng ta có nên ăn mừng cho những tiến bộ về phần cứng máy tính? Còn những tiến bộ trên con đường tiến tới giả lập hoàn chỉnh não bộ thì sao? Chúng ta sẽ lần lượt xem xét hai câu hỏi này.

Những hiệu ứng của tiến bộ phần cứng

Máy tính nhanh hơn sẽ giúp tạo ra trí tuệ máy dễ dàng hơn. Do đó, một hiệu ứng của việc gia tăng tốc độ tiến bộ phần cứng là đẩy nhanh sự xuất hiện của trí tuệ máy. Như đã thảo luận, đây có lẽ là một điều tồi tệ theo quan điểm phi cá nhân vì nó làm giảm lượng thời gian chúng ta có để giải quyết vấn đề kiểm soát và giúp nền văn minh nhân loại trưởng thành hơn. Tuy nhiên, điều này không hẳn đúng. Vì siêu trí tuệ sẽ loại bỏ nhiều rủi ro diệt vong khác, chúng ta có thể có lý do để ưu tiên phát triển sớm hơn nếu các rủi ro diệt vong này là nghiêm trọng.¹⁸

Đẩy nhanh hoặc trì hoãn thời điểm khởi đầu của sự bùng nổ trí tuệ không phải là con đường duy nhất mà qua đó tốc độ tiến bộ phần cứng có thể ảnh hưởng đến mức độ rủi ro diệt vong. Mặt khác, phần cứng có thể thay thế phần mềm ở một mức độ nào đó; vì thế, phần cứng tốt hơn sẽ làm giảm yêu cầu về kỹ năng tối thiểu để mã hóa AI hạt giống. Máy tính nhanh cũng có thể thúc đẩy việc sử dụng các phương pháp tiếp cận phụ thuộc nhiều hơn vào các kỹ thuật xử lý tổng lực (như thuật toán di truyền và các phương pháp kiểu tạo sinh-đánh giá- loại bỏ khác) và ít hơn vào các kỹ thuật đòi hỏi sự hiểu biết sâu sắc khi sử dụng. Nếu các kỹ thuật xử lý tổng lực được áp dụng

vào các thiết kế hệ thống vô chính phủ hoặc thiếu chính xác hơn, trong đó vấn đề kiểm soát khó giải quyết hơn so với các hệ thống được thiết kế chính xác và kiểm soát về mặt lý thuyết, đây sẽ là một cách khác mà máy tính nhanh hơn làm gia tăng mức độ rủi ro diệt vong.

Một vấn đề khác cần cân nhắc là tiến bộ nhanh về phần cứng làm tăng khả năng xảy ra quá trình cất cánh nhanh. Ngành công nghiệp bán dẫn càng phát triển nhanh thì thời gian dành cho lập trình viên để khai thác năng lực của máy tính ở bất kỳ mức hiệu năng nào càng giảm xuống. Điều này đồng nghĩa một sự bùng nổ trí tuệ ít có khả năng khởi phát hơn ở mức hiệu năng phần cứng thấp nhất mà ở đó nó là khả thi. Vì vậy, một sự bùng nổ trí tuệ sẽ có khả năng khởi phát cao hơn khi phần cứng đã tiến bộ vượt quá mức tối thiểu mà tại đó, phương pháp lập trình thành công cuối cùng có thể đã thành công. Sau đó, một mái hiên phần cứng sẽ xuất hiện khi quá trình cất cánh xảy ra. Như chúng ta đã thấy trong Chương 4, mái hiên phần cứng là một trong những yếu tố chính làm giảm sức chống đối trong quá trình cất cánh. Tiến bộ phần cứng nhanh chóng, do đó, sẽ có xu hướng làm cho quá trình chuyển đổi sang siêu trí tuệ nhanh hơn và bùng nổ hơn.

Một quá trình cất cánh nhanh hơn thông qua mái hiên phần cứng có thể ảnh hưởng đến các rủi ro của quá trình chuyển đổi theo nhiều cách. Rõ ràng nhất, cất cánh nhanh hơn mang đến ít cơ hội phản ứng và thực hiện điều chỉnh hơn trong khi quá trình chuyển đổi đang diễn ra, và điều này sẽ có xu hướng tăng rủi ro. Một vấn đề cần xem xét là mái hiên phần cứng sẽ làm giảm khả năng kiềm chế AI hạt giống tự cải thiện nguy hiểm bằng cách hạn chế khả năng chiếm lĩnh

phần cứng của nó: khi một bộ xử lý càng nhanh thì càng cần ít bộ xử lý để AI tự nhảy vọt và trở thành siêu trí tuệ. Tuy nhiên, một hiệu ứng khác của mái hiên phần cứng là cào bằng các dự án lớn và nhỏ bằng cách giảm tầm quan trọng của một trong những lợi thế của các dự án lớn hơn, đó là khả năng mua được máy tính mạnh hơn. Hiệu ứng này cũng có thể làm tăng mức độ rủi ro diệt vong, nếu các dự án lớn hơn có nhiều khả năng thành công hơn trong việc giải quyết vấn đề kiểm soát và theo đuổi các mục tiêu có thể chấp nhận về đạo đức.¹⁹

Quá trình cất cánh nhanh hơn cũng có một số ưu điểm. Cất cánh nhanh hơn sẽ làm tăng khả năng hình thành một thể đơn nhất. Nếu việc thiết lập một thể đơn nhất là đủ quan trọng để giải quyết các vấn đề phối hợp hậu chuyển đổi, chúng ta có thể chấp nhận rủi ro lớn hơn trong khi xảy ra sự bùng nổ trí tuệ nhằm giảm thiểu rủi ro xảy ra những thất bại nghiêm trọng về phối hợp hậu bùng nổ.

Sự phát triển trong điện toán có thể ảnh hưởng đến kết quả của một cuộc cách mạng trí tuệ máy không chỉ bằng cách đóng vai trò trực tiếp trong việc xây dựng trí tuệ máy mà còn có tác động lan tỏa đến xã hội, gián tiếp giúp hình thành các điều kiện ban đầu của sự bùng nổ trí tuệ. Với yêu cầu phần cứng đủ tốt, cho phép máy tính cá nhân được sản xuất hàng loạt với chi phí thấp, Internet hiện đang ảnh hưởng đến hoạt động của con người trong nhiều lĩnh vực, bao gồm cả nghiên cứu về trí tuệ nhân tạo và vấn đề kiểm soát. (Nếu không có Internet, cuốn sách này có thể chưa được viết hoặc bạn sẽ không tìm thấy nó.) Tuy nhiên, phần cứng hiện đã đủ tốt để sử dụng rất nhiều ứng dụng hỗ trợ giao tiếp và thảo luận của con người, và việc tốc độ tiến bộ của các lĩnh vực này bị kìm hãm bởi tốc độ cải thiện phần cứng là không rõ ràng.²⁰

Sau khi cân nhắc, xem ra tiến bộ nhanh hơn về phần cứng máy tính lại là điều không mong muốn từ quan điểm đánh giá phi cá nhân. Kết luận tạm thời này có thể bị lật đổ nếu, chẳng hạn, các mối đe dọa từ các rủi ro diệt vong khác hoặc từ các thất bại trong phối hợp hậu chuyển đổi trở nên vô cùng lớn. Trong mọi trường hợp, dường như khó có nhiều động lực cho tốc độ tiến bộ phần cứng. Do đó, những nỗ lực của chúng ta nhằm cải thiện các điều kiện ban đầu của sự bùng nổ trí tuệ có lẽ nên tập trung vào các tham số khác.

Lưu ý rằng ngay cả khi chúng ta không thể tìm ra cách ảnh hưởng đến một số tham số thì việc xác định “dấu” của nó (có nghĩa là chúng ta muốn tham số đó tăng hay giảm) vẫn có thể hữu dụng như bước đi đầu tiên trong việc lập kế hoạch chiến lược. Sau này, chúng ta có thể khám phá một “điểm tựa” mới cho phép thao tác với tham số này dễ dàng hơn. Hoặc, chúng ta cũng có thể phát hiện ra rằng dấu của tham số này có liên quan đến dấu của một tham số khác dễ thao tác hơn, sao cho những phân tích ban đầu của chúng ta có thể giúp ích trong việc quyết định điều cần làm với tham số khác đó.

Có nên thúc đẩy nghiên cứu giả lập hoàn chỉnh não bộ?

Việc giải quyết vấn đề kiểm soát trí tuệ nhân tạo càng khó khăn thì việc thúc đẩy con đường giả lập hoàn chỉnh não bộ như một phương án thay thế càng trở nên cấp thiết. Tuy nhiên, có một số vấn đề phải được phân tích trước khi chúng ta có thể đưa ra một quyết định được xem xét kỹ lưỡng.

Đầu tiên, chúng ta có vấn đề về khớp nối công nghệ, đã được thảo luận ở trên. Chúng ta đã chỉ ra rằng nỗ lực phát triển giả lập

hoàn chỉnh não bộ thay vào đó lại có thể tạo ra AI hình thái thần kinh, một dạng trí tuệ máy đặc biệt nguy hiểm.

Nhưng để lập luận, hãy giả sử chúng ta thực sự thành công trong với giả lập hoàn chỉnh não bộ (WBE). Liệu nó có an toàn hơn AI? Bản thân điều này cũng là một vấn đề phức tạp. Có ít nhất ba lợi thế giả định của WBE: (i) các đặc điểm hiệu năng của nó sẽ được hiểu rõ hơn so với AI; (ii) nó sẽ kế thừa động lực của con người; và (iii) nó sẽ dẫn đến một quá trình cất cánh chậm hơn. Hãy cùng suy nghĩ rất ngắn gọn về mỗi lợi thế.

i. Sẽ hợp lý nếu nói rằng các đặc điểm hiệu năng trí tuệ của một bản giả lập là dễ hiểu hơn so với một AI. Chúng ta có nhiều kinh nghiệm với những điểm mạnh và điểm yếu của trí tuệ con người nhưng không có kinh nghiệm với trí tuệ nhân tạo cấp độ con người. Tuy nhiên, hiểu những gì mà trí tuệ con người đã số hóa có thể và không thể làm không giống như việc hiểu được trí tuệ này sẽ phản ứng ra sao với những sửa đổi nhằm tăng cường hiệu năng cho nó. Một trí tuệ nhân tạo, ngược lại, có thể được thiết kế cẩn thận để trở nên dễ hiểu trong cả hai trạng thái tĩnh và động. Vì vậy, trong khi giả lập hoàn chỉnh não bộ có thể dễ dự đoán hơn về hiệu năng trí tuệ so với một AI đặc trưng ở giai đoạn phát triển tương đương, chúng ta vẫn chưa biết rõ liệu giả lập hoàn chỉnh não bộ có thể dễ dự đoán hơn ở trạng thái động so với một AI được lập trình bởi các lập trình viên có ý thức về vấn đề an toàn.

ii. Đối với một bản giả lập kế thừa các động lực của người làm mẫu, điều này không được đảm bảo. Việc nắm bắt các xu hướng mang tính đánh giá của con người có thể đòi hỏi một bản giả lập có độ trung thực rất cao. Ngay cả khi một số động lực cá nhân đã được

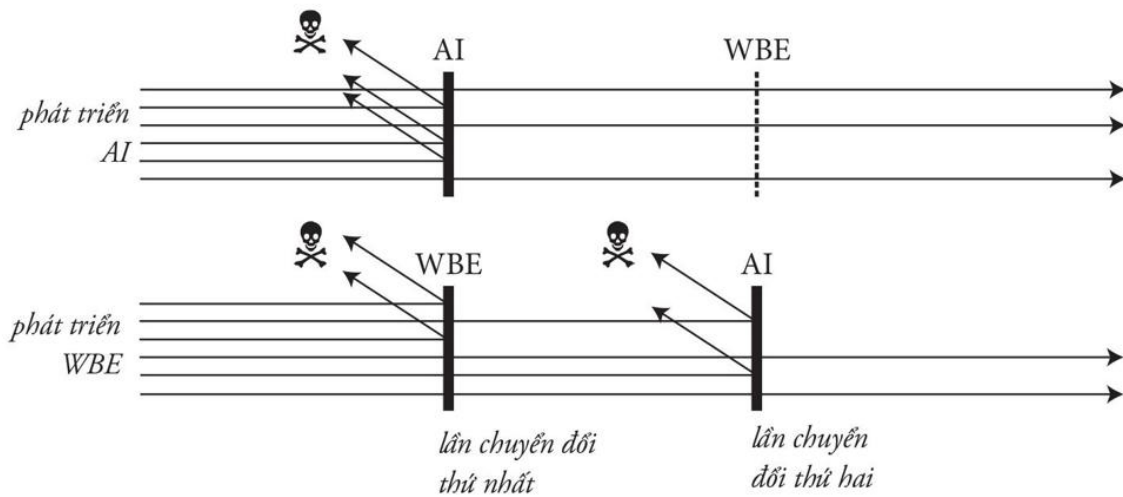
nắm bắt một cách hoàn hảo thì mức an toàn có được vẫn là không rõ ràng. Con người có thể không đáng tin, ích kỷ và tàn nhẫn. Mặc dù người làm mẫu được lựa chọn vì tư cách xuất chúng, nhưng khó có thể thấy trước ai đó sẽ hành động như thế nào khi được đưa vào một hoàn cảnh hoàn toàn lạ lẫm, được tăng cường trí tuệ siêu phàm và bị cám dỗ bởi cơ hội thống trị thế giới. Đúng là các bản giả lập ít nhất sẽ có khả năng sở hữu động lực như của con người (chứ không chỉ đánh giá cao việc làm kẹp giấy hoặc khám phá các chữ số của pi), nhưng điều này sẽ có hoặc không thể đảm bảo, tùy quan điểm cá nhân về bản chất con người.²²

iii. Chúng ta vẫn chưa rõ tại sao giả lập hoàn chỉnh não bộ dẫn đến quá trình cất cánh chậm hơn so với trí tuệ nhân tạo. Có lẽ với giả lập hoàn chỉnh não bộ, người ta sẽ mong đợi một mái hiên phần cứng nhỏ hơn vì kỹ thuật này ít hiệu quả hơn so với AI. Có lẽ một hệ thống AI cũng có thể hấp thụ dễ dàng hơn toàn bộ công suất tính toán có sẵn vào một trí tuệ tích hợp khổng lồ, trong khi đó, giả lập hoàn chỉnh não bộ sẽ chỉ vượt qua được trí tuệ siêu chất lượng và nhân loại về tốc độ cùng số lượng cá thể. Nếu giả lập hoàn chỉnh não bộ dẫn đến một quá trình cất cánh chậm hơn, điều này có thể mang lại lợi ích trong việc giải quyết vấn đề kiểm soát. Cất cánh chậm hơn cũng sẽ khiến một kết quả đa cực có nhiều khả năng xảy ra hơn. Tuy nhiên, việc một kết quả đa cực có phải là điều mong muốn hay không vẫn là rất đáng nghi ngờ.

Với ý tưởng cho rằng sẽ an toàn hơn nếu phát triển giả lập hoàn chỉnh não bộ trước, một vấn đề phức tạp khác cũng quan trọng không kém, đó là: sự cần thiết của việc đối phó với một lần chuyển đổi thứ hai. Ngay cả khi dạng thức trí tuệ máy cấp độ con người đầu

tiên là dựa trên giả lập, trí tuệ nhân tạo vẫn có thể được phát triển. AI ở dạng trưởng thành có những lợi thế quan trọng so với WBE, khiến nó cuối cùng sẽ trở thành công nghệ mạnh mẽ hơn.²³ AI trưởng thành sẽ khiến WBE trở nên lỗi thời (ngoại trừ với mục đích đặc biệt là bảo lưu tâm trí con người), trong khi điều ngược lại là không thể.

Điều này nghĩa là nếu AI được phát triển trước, sự bùng nổ trí tuệ có thể sẽ chỉ xảy ra một lần. Nhưng nếu WBE được phát triển đầu tiên, có thể sẽ có hai đợt: sự xuất hiện của WBE; và sau đó là sự xuất hiện của AI. Tổng lượng rủi ro diệt vong của con đường phát triển WBE sẽ là tổng lượng rủi ro của hai lần chuyển đổi (với điều kiện chúng ta đã vượt qua lần chuyển đổi đầu tiên); xem Hình 13.²⁴



Hình 13: Phát triển trí tuệ nhân tạo hay giả lập hoàn chỉnh não bộ trước? Khi phát triển AI trước, chỉ có một lần chuyển đổi tạo ra rủi ro diệt vong. Khi phát triển WBE trước, ta sẽ có hai lần chuyển đổi đầy rủi ro, đầu tiên là sự phát triển của WBE và sau đó là sự phát triển của AI. Tổng lượng rủi ro diệt vong ở kịch bản phát triển WBE là tổng

của các rủi ro này. Tuy nhiên, nguy cơ chuyển đổi AI có thể thấp hơn nếu xảy ra ở một thế giới mà WBE đã được thực hiện thành công.

Quá trình chuyển đổi AI sẽ an toàn hơn bao nhiêu trong một thế giới WBE? Có thể quá trình chuyển đổi AI sẽ ít bùng nổ hơn nếu nó xảy ra sau khi một số dạng thức trí tuệ máy đã được chế tạo. Các bản giả lập chạy ở tốc độ kỹ thuật số và có số lượng vượt xa con người sẽ làm giảm sự khác biệt về nhận thức, giúp chúng dễ dàng kiểm soát AI hơn. Viễn cảnh này không quá nặng nề, vì khoảng cách giữa AI và WBE có thể vẫn lớn. Tuy nhiên, nếu các bản giả lập không chỉ nhanh hơn, nhiều hơn mà còn thông minh hơn về chất so với con người (hoặc ít nhất là tương đương những người thông minh nhất) thì kịch bản phát triển WBE sẽ có lợi thế tương đương việc tăng cường nhận thức của con người, mà chúng ta đã thảo luận ở trên.

Một vấn đề khác cần xem xét là việc chuyển đổi sang WBE sẽ gia tăng khoảng cách dẫn đầu của dự án đi trước. Hãy xem xét một kịch bản mà người đi trước có thời gian dẫn trước sáu tháng so với dự án bám đuổi gần nhất trong việc phát triển công nghệ giả lập hoàn chỉnh não bộ. Giả sử các bản giả lập đầu tiên được tạo ra có tính hợp tác, kiên nhẫn và tập trung vào mức độ an toàn. Nếu chạy trên phần cứng có tốc độ cao, các bản giả lập này có thể tiêu tốn hàng thiên niên kỷ để suy nghĩ cách tạo ra AI an toàn. Ví dụ: nếu chạy ở tốc độ gấp 100.000 lần và có thể nghiên cứu vấn đề kiểm soát mà không bị xao lãng trong sáu tháng theo thời gian thiên văn, chúng có thể giải quyết vấn đề này trong 50 thiên niên kỷ trước khi phải đối mặt với sự cạnh tranh từ các bản giả lập khác. Nếu có phần cứng phù hợp, chúng có thể đẩy nhanh sự tiến bộ của mình bằng cách tạo ra

vô số bản sao để hoạt động độc lập trên các vấn đề nhỏ. Nếu dự án đi trước sử dụng khoảng thời gian sáu tháng ấy để thiết lập một thể đơn nhất, họ có thể kéo dài cho nhóm phát triển AI giả lập của mình một lượng thời gian không giới hạn để giải quyết vấn đề kiểm soát.²⁵

Sau khi cân nhắc, có vẻ như nguy cơ chuyển đổi AI sẽ giảm nếu WBE xuất hiện trước AI. Tuy nhiên, khi chúng ta kết hợp rủi ro còn sót lại trong quá trình chuyển đổi AI với rủi ro của quá trình chuyển đổi WBE trước đó, mức độ rủi ro diệt vong của con đường phát triển WBE ra sao khi so với rủi ro của con đường phát triển AI sẽ là một vấn đề rất không rõ ràng. Chỉ khi chúng ta bi quan về khả năng của con người trong việc đối phó với một quá trình chuyển đổi AI (khi đã tính đến việc bản chất hoặc nền văn minh của loài người có thể đã được cải thiện khi chúng ta đối mặt với thử thách này) thì con đường phát triển WBE trước xem ra mới hấp dẫn.

Để biết liệu công nghệ giả lập hoàn chỉnh não bộ có nên được thúc đẩy hay không, có một số điểm quan trọng hơn nữa cần đưa vào xem xét. Trong đó, đáng kể nhất là khớp nối công nghệ được đề cập trước đó: việc đẩy mạnh tốc độ phát triển WBE thay vào đó có thể tạo ra AI hình thái thần kinh. Đây là một lý do chống lại việc thúc đẩy WBE.²⁶ Không còn nghi ngờ gì nữa, có một số thiết kế AI tổng hợp kém an toàn hơn so với một số thiết kế hình thái thần kinh. Tuy nhiên, theo dự kiến, các thiết kế hình thái thần kinh xem ra vẫn nguy hiểm hơn. Một lý do cho điều này là việc bắt chước có thể thay thế hiểu biết. Để xây dựng một hệ thống từ nền móng, người ta thường phải hiểu tương đối rõ về cách thức hoạt động của nó, nhưng hiểu biết như vậy có thể không cần thiết khi chỉ cần sao chép các tính năng của một hệ thống hiện có. Giả lập hoàn chỉnh não bộ dựa trên

quá trình sao chép tổng thể từ sinh học, một việc có thể không đòi hỏi sự hiểu biết toàn diện về nhận thức ở cấp độ hệ thống tính toán (mặc dù chắc chắn sẽ cần một lượng lớn hiểu biết ở cấp độ thành phần). Về khía cạnh này, AI hình thái thần kinh có thể giống với giả lập hoàn chỉnh não bộ: chúng ta có thể có được dạng AI này bằng cách ghép các mảnh ghép sao chép từ sinh học mà không cần các kỹ sư phải hiểu biết sâu sắc về cách thức hoạt động của hệ thống. Nhưng AI hình thái thần kinh sẽ khác giả lập hoàn chỉnh não bộ ở chỗ: nó không có động lực của con người theo mặc định.²⁷ Đây là lập luận phản đối việc theo đuổi phương pháp giả lập hoàn chỉnh não bộ đến mức độ có thể tạo ra AI hình thái thần kinh.

Vấn đề thứ hai cần cân nhắc là WBE có nhiều khả năng sẽ thông báo trước cho chúng ta về sự xuất hiện của nó. Với AI, một ai đó sẽ luôn có thể tạo ra một bước đột phá bất ngờ về mặt khái niệm. Ngược lại, WBE sẽ đòi hỏi nhiều bước đi tiên quyết tốn nhiều công sức (như các phương tiện quét công suất cao, phần mềm xử lý hình ảnh, nghiên cứu mô hình thần kinh chi tiết). Do đó, chúng ta có thể tin tưởng rằng WBE không phải gần kề (giả sử, không ít hơn 15-20 năm nữa). Điều này có nghĩa là những nỗ lực tăng tốc WBE sẽ tạo ra sự khác biệt chủ yếu trong các kịch bản mà trí tuệ máy được phát triển tương đối muộn. Điều này có thể khiến các khoản đầu tư vào WBE trở nên hấp dẫn với những người muốn sự bùng nổ trí tuệ loại bỏ các rủi ro diệt vong khác, nhưng lại cảnh giác với việc hỗ trợ AI vì sợ rằng sự bùng nổ này sẽ xảy ra sớm hơn, trước khi vấn đề kiểm soát được giải quyết. Tuy nhiên, sự không chắc chắn trong khoảng thời gian có liên quan hiện nay có thể vẫn là quá lớn để cho phép sự cân nhắc này có trọng lượng đáng kể.²⁸

Vì thế, một chiến lược thúc đẩy WBE sẽ là hấp dẫn nhất nếu (a) chúng ta rất bi quan về con người giải quyết vấn đề kiểm soát trên AI, (b) chúng ta không quá lo lắng về AI hình thái thần kinh, kết quả đa cực hoặc rủi ro xảy ra lần chuyển đổi thứ hai, (c) chúng ta cho rằng thời điểm mặc định của WBE và AI đã gần kề, và (d) chúng ta muốn siêu trí tuệ được phát triển không quá muộn cũng không quá sớm.

Ưu tiên tốc độ từ quan điểm ảnh hưởng đến cá nhân

Tôi sợ rằng một người bình luận blog nào đó với nickname “washbash” có thể đại diện cho nhiều người khi viết:

Theo bản năng, tôi sẽ chọn tăng tốc. Không phải vì tôi nghĩ rằng điều này là tốt hơn cho thế giới. Tại sao tôi phải quan tâm đến thế giới khi mà mình đã chết? Tôi chỉ muốn nó đến nhanh thôi, chết tiệt! Điều này giúp tôi có thêm cơ hội trải nghiệm một tương lai với công nghệ hiện đại hơn.²⁹

Từ quan điểm ảnh hưởng đến cá nhân, chúng ta có lý do lớn hơn để tăng tốc với toàn bộ các loại công nghệ cao có thể gây ra rủi ro diệt vong. Lý do của việc này nằm ở kết quả mặc định rằng hầu hết mọi người sẽ chết trong một thế kỷ tới.

Việc đẩy nhanh tốc độ là đặc biệt hiệu quả đối với các công nghệ kéo dài sự sống, và do đó làm gia tăng số lượng người vẫn còn tồn tại khi xảy ra sự bùng nổ trí tuệ. Nếu cuộc cách mạng trí tuệ diễn ra suôn sẻ, siêu trí tuệ được tạo ra gần như chắc chắn có thể nghĩ ra cách kéo dài vĩnh viễn tuổi thọ của những ai còn sống. Nó không chỉ duy trì sự sống cho con người mà còn phục hồi sức khỏe và sự trẻ trung năng động, và nâng cao năng lực vượt xa suy nghĩ hiện nay của chúng ta về giới hạn của con người; hoặc giúp con người “thoát xác hoàn hồn” bằng cách tải tâm trí lên một chất nền kỹ thuật số và

ban cho những linh hồn đã được “phóng thích” này những thân thể ảo hoàn mỹ. Đối với các công nghệ không có tính “cứu rỗi” như vậy, động lực đẩy nhanh tốc độ sẽ yếu hơn, nhưng có lẽ vẫn sẽ được ủng hộ với hy vọng nâng cao chất lượng cuộc sống.³⁰

Cùng một lối lý luận sẽ khiến quan điểm ảnh hưởng đến cá nhân ủng hộ nhiều sự đổi mới công nghệ đầy rủi ro, hứa hẹn thúc đẩy sự bùng nổ trí tuệ, ngay cả khi những đổi mới này bị bóp méo trong một viễn cảnh phi cá nhân. Những đổi mới như vậy có thể rút ngắn những thời khắc tăm tối mà khi đó cá nhân chúng ta phải bám víu vào những “cành cong” để sống sót và chứng kiến buổi bình minh của thời đại hậu nhân. Từ quan điểm ảnh hưởng đến cá nhân, tiến bộ phần cứng nhanh hơn xem ra là điều đáng mong đợi, cũng giống như bước tiến tới WBE. Bất kỳ tác động bất lợi nào đối với mức độ rủi ro diệt vong có lẽ đều bị lấn át bởi lợi ích cá nhân của việc gia tăng khả năng bùng nổ trí tuệ ngay trong thời đại của những con người đang sống.³¹

Hợp tác

Mức độ mà thế giới có thể phối hợp và hợp tác trong việc phát triển trí tuệ máy là một tham số quan trọng. Hợp tác sẽ mang lại nhiều lợi ích. Chúng ta hãy xem tham số này ảnh hưởng đến kết quả ra sao, và liệu chúng ta có những đòn bẩy nào để gia tăng quy mô và mức độ của việc hợp tác này.

Động lực chạy đua và những mối nguy hiểm

Một cuộc chạy đua động lực tồn tại khi một dự án lo sợ bị một dự án khác qua mặt. Điều này không đòi hỏi sự tồn tại trên thực tế của nhiều dự án. Trong tình huống chỉ có một dự án, một cuộc chạy đua động lực vẫn có thể diễn ra nếu dự án đó không biết về việc mình

đang thiếu đối thủ cạnh tranh. Quân Đồng minh có thể đã không phát triển bom nguyên tử nhanh như vậy nếu không (nhằm lẫn) tin rằng người Đức có thể tiến gần tới mục tiêu này.

Mức độ nghiêm trọng của một động lực chạy đua (mức độ ưu tiên dành cho tốc độ so với độ an toàn của các đối thủ cạnh tranh) phụ thuộc vào một số yếu tố, chẳng hạn như sự sát sao của cuộc đua, tương quan về tầm quan trọng giữa năng lực và may mắn, số lượng đối thủ, mức độ khác nhau của các phương pháp tiếp cận và mức độ chia sẻ mục tiêu của các dự án. Niềm tin của các đối thủ cạnh tranh vào các yếu tố này cũng có liên quan. (Xem Khung 13.)

Khung 13 Cuộc chạy đua rủi ro xuống đáy

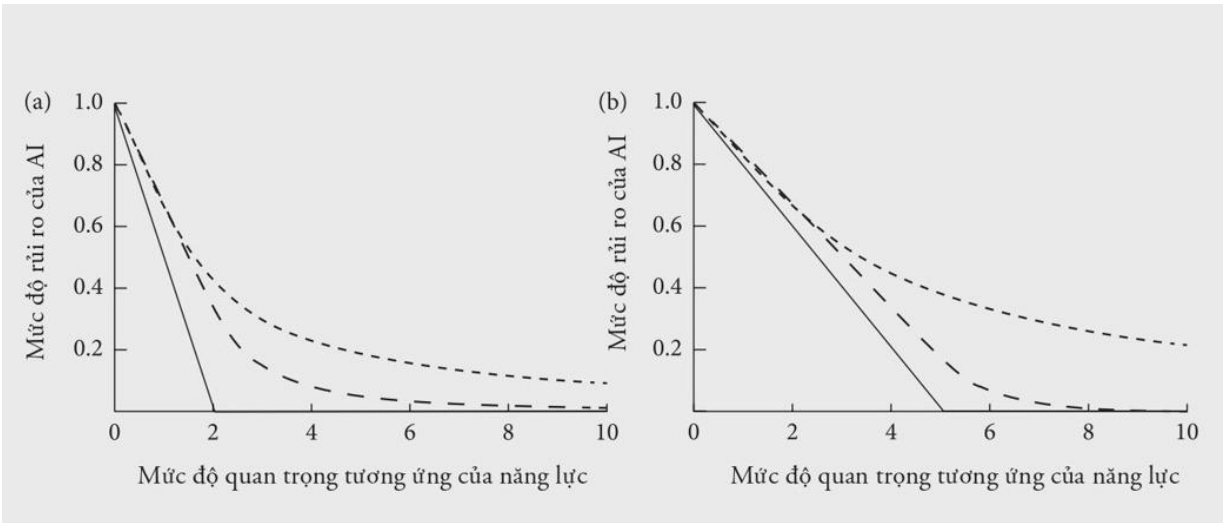
Hãy xem xét một cuộc chạy đua vũ trang AI giả định, trong đó có một số nhóm cạnh tranh để phát triển siêu trí tuệ.³² Mỗi nhóm tự quyết định lượng đầu tư vào mức độ an toàn, khi biết rằng khoản tài nguyên dành cho việc phát triển các biện pháp phòng ngừa an toàn sẽ không dành cho việc phát triển AI. Khi không có thỏa thuận giữa tất cả các đối thủ cạnh tranh (có thể bị cản trở bởi những khó khăn trong việc đàm phán hoặc thực thi), sẽ dễ dẫn đến một cuộc “chạy đua rủi ro xuống đáy” (risk-race to the bottom), khiến mỗi nhóm đều chỉ thực hiện các biện pháp phòng ngừa tối thiểu.

Ai đó có thể mô hình hóa hiệu năng của từng nhóm như một hàm năng lực (đo thô khả năng và may mắn) của nhóm đó và một thừa số phạt tương ứng với chi phí phòng ngừa an toàn. Nhóm có hiệu năng cao nhất tạo ra AI đầu tiên. Rủi ro của AI đó được xác định bởi lượng đầu tư của người tạo ra nó vào sự an toàn. Trong trường hợp xấu nhất, tất cả các nhóm sẽ có mức độ năng lực như nhau. Nhóm chiến thắng sau đó sẽ được xác định riêng bằng khoản đầu tư vào sự an

toàn: nhóm nào thực hiện ít biện pháp phòng ngừa an toàn nhất sẽ chiến thắng. Điểm cân bằng Nash cho cuộc đua này là điểm mà tại đó toàn bộ các nhóm không đầu tư cho mức độ an toàn. Trong thế giới thực, một tình huống như vậy có thể xảy ra thông qua một guồng máy rủi ro (risk racket): một nhóm nào đó, vì sợ bị tụt lại phía sau, sẽ gia tăng khả năng chấp nhận rủi ro để bắt kịp với các đối thủ cạnh tranh của mình, và các nhóm khác cũng phản ứng theo cách như vậy cho đến khi đạt mức rủi ro cực đại.

Năng lực và rủi ro

Tình hình thay đổi khi năng lực trở nên biến hóa. Khi mức năng lực trở nên quan trọng hơn so với chi phí phòng ngừa an toàn, guồng máy rủi ro sẽ suy yếu: các nhóm sẽ có ít động lực hơn để chịu thêm rủi ro nếu làm như vậy không có khả năng thay đổi thứ tự của cuộc đua. Điều này được minh họa qua các kịch bản khác nhau trong Hình 14, cho thấy mức độ rủi ro của AI phụ thuộc vào tầm quan trọng của năng lực. Lượng đầu tư vào sự an toàn nằm trong khoảng từ 1 (dẫn đến AI hoàn toàn an toàn) đến 0 (AI hoàn toàn không an toàn). Trục x biểu thị tầm quan trọng tương đối của năng lực so với đầu tư an toàn trong việc xác định tốc độ tiến bộ của nhóm hướng đến AI. (Ở mức 0,5, mức đầu tư an toàn quan trọng gấp đôi năng lực; ở mức 1 là bằng nhau; ở mức 2, năng lực quan trọng gấp đôi an toàn;...) Trục y biểu thị mức độ rủi ro AI (tỷ lệ dự kiến so với độ thỏa dụng tối đa mà nhóm chiến thắng có được).



Hình 14: Mức độ rủi ro trong các cuộc đua công nghệ AI. Mức độ rủi ro của AI nguy hiểm trong sơ đồ đơn giản của một cuộc đua công nghệ liên quan đến (a) hai nhóm hoặc (b) năm nhóm, được vẽ theo mức độ quan trọng tương đối của năng lực (ngược với mức đầu tư vào sự an toàn) trong việc xác định dự án nào chiến thắng cuộc đua. Các biểu đồ hiển thị ba kịch bản về mức độ thông tin: không có thông tin về năng lực (nét liền), thông tin về năng lực được bảo mật (nét đứt) và thông tin về năng lực đầy đủ (đường chấm).

Chúng ta thấy rằng, trong tất cả các kịch bản, mức độ nguy hiểm của AI được tạo ra là tối đa khi năng lực không có vai trò, và giảm dần khi tầm quan trọng của năng lực tăng lên.

Các mục tiêu tương thích

Một cách khác để giảm rủi ro là để các nhóm góp phần nhiều hơn vào thành công của nhau. Nếu các đối thủ cạnh tranh bị thuyết phục rằng về thứ hai đồng nghĩa với việc mất đi mọi thứ họ quan tâm, họ sẽ chấp nhận mọi rủi ro cần thiết để vượt qua đối thủ của mình. Ngược lại, các đội sẽ đầu tư nhiều hơn vào sự an toàn nếu ít bị lệ

thuộc vào việc chiến thắng cuộc đua. Điều này cho thấy chúng ta nên khuyến khích các hình thức đầu tư chéo.

Số lượng đối thủ cạnh tranh

Số lượng các nhóm cạnh tranh càng nhiều, cuộc đua càng trở nên nguy hiểm: mỗi đội, với ít cơ hội chiến thắng hơn, sẽ sẵn sàng vứt bỏ sự thận trọng hơn. Điều này có thể được nhìn thấy bằng cách đối chiếu Hình 14a (hai nhóm) với Hình 14b (năm nhóm). Trong mọi kịch bản, nhiều đối thủ cạnh tranh hơn đồng nghĩa với rủi ro nhiều hơn. Rủi ro sẽ giảm nếu các nhóm kết hợp thành một số lượng nhỏ hơn các liên minh cạnh tranh.

“Lời nguyên” dư thừa thông tin

Liệu có ổn nếu các nhóm biết vị trí của mình trong cuộc đua (chẳng hạn như biết được năng lực của mình đạt bao nhiêu điểm)? Ở đây, các yếu tố đối nghịch đang tạo ra ảnh hưởng. Chúng ta mong đợi nhóm dẫn đầu biết mình đang dẫn đầu (để nhóm này biết rằng mình có thêm cơ hội áp dụng các biện pháp an toàn). Tuy nhiên, điều không mong muốn là nhóm bị tụt lại phía sau biết được vị trí của mình (vì nhóm này sẽ phải phải cắt giảm khoản đầu tư vào sự an toàn để có hy vọng bắt kịp đối thủ). Tuy theo trực giác có vẻ như cán cân có thể nghiêng về bất kỳ bên nào, nhưng các đồ thị lại rất rõ ràng: thông tin (theo dự kiến) là xấu.³³ Mỗi hình 14a và 14b đưa ra ba kịch bản: các đường nét liền tương ứng với các tình huống mà trong đó không nhóm nào biết điểm số về năng lực, kể cả của bản thân. Các đường nét đứt cho thấy các tình huống trong đó mỗi nhóm đều chỉ biết khả năng của riêng mình. (Trong những tình huống đó, một nhóm chỉ chịu rủi ro cao hơn nếu có năng lực thấp.) Và các đường chấm cho thấy điều gì xảy ra khi tất cả các nhóm biết năng lực của

nhau. (Họ chấp nhận rủi ro thêm nếu điểm khả năng của họ suy tsoát nhau.) Khi các đội càng có thêm thông tin, động lực của cuộc đua càng trở nên tồi tệ hơn.

Trong quá trình phát triển siêu trí tuệ máy, xem ra sẽ có ít nhất một động lực chạy đua mức độ nhẹ, và một động lực chạy đua mạnh mẽ cũng là khả thể. Động lực chạy đua có hệ quả quan trọng đối với cách chúng ta nên nghĩ về thách thức chiến lược được đặt ra bởi khả năng xảy ra một sự bùng nổ trí tuệ.

Động lực chạy đua có thể thúc đẩy các dự án tiến nhanh hơn tới siêu trí tuệ đồng, thời giảm đầu tư vào việc giải quyết vấn đề kiểm soát. Các tác động bất lợi khác của động lực chạy đua cũng có thể xảy ra, chẳng hạn như sự đối đầu trực tiếp giữa các đối thủ cạnh tranh. Giả sử hai quốc gia đang chạy đua xem ai phát triển siêu trí tuệ trước, và một trong số họ đang dẫn trước. Trong một tình huống “được ăn cả”, một dự án tụt hậu có thể sẽ có động lực khởi động một cuộc tấn công tuyệt vọng chống lại đối thủ hơn là thụ động chờ đợi thất bại. Dự đoán được khả năng này, người dẫn trước có thể sẽ muốn tấn công phủ đầu. Nếu “nhân vật phản diện” là những quốc gia hùng mạnh, cuộc đụng độ có thể sẽ rất đẫm máu.³⁴ (Một cuộc “tấn công phẫu thuật” nhắm vào dự án AI của đối thủ có thể có nguy cơ dẫn đến một cuộc đối đầu quy mô lớn hơn và trong bất kỳ trường hợp nào, có thể không khả thi nếu quốc gia sở hữu dự án đã có phòng bị.³⁵)

Kịch bản trong đó các đối thủ không phải là các quốc gia mà là các thể chế nhỏ hơn, chẳng hạn như phòng thí nghiệm của công ty hoặc nhóm học thuật, có thể sẽ ít mang tính tàn phá hơn khi xảy ra mâu thuẫn. Tuy nhiên, hậu quả cuối cùng của việc cạnh tranh sẽ tồi

tệ không kém. Điều này là tác hại dự kiến của việc sự cạnh tranh không chủ yếu bắt nguồn từ sự tàn phá của cuộc chiến mà từ sự suy giảm của các biện pháp phòng ngừa. Một động lực chạy đua, như chúng ta đã thấy, làm giảm lượng đầu tư vào sự an toàn; và mâu thuẫn, ngay cả khi bất bạo động, cũng sẽ có xu hướng chặn đứng các cơ hội hợp tác, vì các dự án sẽ ít có khả năng hơn trong việc chia sẻ ý tưởng để giải quyết vấn đề kiểm soát trong môi trường thù địch và ngờ vực.³⁶

Về lợi ích của việc hợp tác

Hợp tác có rất nhiều lợi ích. Nó làm giảm sự vội vàng trong việc phát triển trí tuệ máy, cho phép đầu tư lớn hơn vào sự an toàn, giúp tránh được các hành vi bạo lực, và tạo điều kiện cho việc chia sẻ ý tưởng về cách giải quyết vấn đề kiểm soát. Chúng ta cũng có thể bổ sung thêm: sự hợp tác sẽ có xu hướng tạo ra kết quả mà trong đó, những thành quả của một sự bùng nổ trí tuệ được kiểm soát thành công được phân phối đều hơn.

Hợp tác rộng rãi hơn dẫn đến việc chia sẻ lợi nhuận rộng rãi hơn không phải là một tiên đề. Về nguyên tắc, một dự án nhỏ được điều hành bởi một người vị tha có thể dẫn đến một kết quả mà trong đó, lợi ích được chia sẻ đồng đều hoặc công bằng giữa tất cả các cá thể đáng cân nhắc về đạo đức. Tuy nhiên, có một số lý do để cho rằng sự hợp tác rộng hơn, liên quan đến số lượng các nhà tài trợ nhiều hơn, (theo dự kiến) là vượt trội về mặt phân phối. Một lý do trong số đó là các nhà tài trợ có lẽ thích một kết quả mà bản thân họ nhận được (ít nhất là) những gì mình xứng đáng. Một sự hợp tác rộng rãi đồng nghĩa với có tương đối nhiều cá nhân nhận được ít nhất là phần mà họ xứng đáng nếu dự án thành công. Một lý do khác là sự

hợp tác rộng rãi xem ra cũng có nhiều khả năng mang lại lợi ích cho người không tham gia hợp tác. Một sự hợp tác rộng rãi hơn sẽ có nhiều người tham gia hơn, vì vậy những người trong cuộc tìm kiếm lợi ích từ bên ngoài sẽ tạo ra sự ràng buộc với nhiều người ngoài cuộc hơn. Một sự hợp tác rộng rãi hơn cũng sẽ có nhiều khả năng sở hữu những đối tượng tham gia vị tha và muốn đem lại lợi ích cho mọi người hơn. Hơn nữa, sự hợp tác rộng rãi hơn có nhiều khả năng hoạt động dưới sự giám sát của công chúng, làm giảm nguy cơ “chiếc bánh” bị thâm tóm bởi một nhóm nhỏ các lập trình viên hoặc nhà đầu tư tư nhân.³⁷ Lưu ý rằng hợp tác thành công càng lớn thì chi phí phải bỏ ra cho việc mở rộng lợi ích cho tất cả những người ngoài cuộc càng thấp. (Ví dụ: nếu 90% mọi người đã tham gia hợp tác, họ sẽ phải trả không quá 10% những gì mình có để mang lại lợi ích tương đương cho những ai không tham gia.)

Do đó, có thể sự hợp tác rộng rãi hơn sẽ có xu hướng dẫn đến phân phối lợi nhuận rộng hơn (mặc dù một số dự án ít nhà tài trợ cũng có thể có những mục tiêu lý tưởng về mặt phân phối). Nhưng tại sao một sự phân phối rộng rãi của lợi ích lại là điều đáng mong đợi?

Có cả lập luận về đạo đức (moral reason) và lập luận cân nhắc lợi-hại (prudential reason) để ủng hộ kết quả mà ở đó mọi người đều nhận được một phần lợi ích. Chúng tôi sẽ không nói nhiều về lập luận đạo đức, ngoại trừ việc lưu ý rằng nó không cần phải dựa trên bất kỳ nguyên tắc bình đẳng nào. Lập luận này có thể dựa trên vấn đề công bằng. Một dự án tạo ra siêu trí tuệ máy sẽ đặt ra một ngoại ứng rủi ro toàn cầu. Mọi người trên hành tinh đều gặp nguy hiểm, bao gồm cả những người không đồng ý để cuộc sống riêng của họ

và những người trong gia đình bị đe dọa như vậy. Vì mọi người đều phải hứng chịu rủi ro nên xem ra việc mọi người đều có được lợi ích sẽ là một yêu cầu công bằng tối thiểu.

Việc tổng lợi ích (dự kiến) có vẻ lớn hơn trong các kịch bản hợp tác là một lý do quan trọng khác khiến các kịch bản như vậy được ưa chuộng hơn về đạo đức.

Lập luận cân nhắc lợi-hại để ủng hộ việc phân phối lợi nhuận rộng có hai phần. Phần thứ nhất là phân phối rộng rãi cần thúc đẩy sự hợp tác, từ đó giảm thiểu hậu quả tiêu cực của động lực chạy đua. Sẽ có ít động lực cạnh tranh nhằm giành vị trí dẫn đầu trong quá trình xây dựng siêu trí tuệ đầu tiên nếu mọi người được lợi như nhau từ thành công của bất kỳ dự án nào. Các nhà tài trợ của một dự án cụ thể cũng có thể được hưởng lợi từ việc đưa ra một tín hiệu đáng tin cậy của một cam kết về việc phân phối đều các lợi ích, một dự án có tính vị tha có nhiều khả năng thu hút thêm nhà tài trợ và giảm số lượng kẻ thù.³⁸

Các khía cạnh khác của lập luận cân nhắc lợi-hại cho việc ủng hộ phân phối lợi nhuận rộng rãi có liên quan đến việc liệu các tác tử sẽ không thích rủi ro hay có các hàm thỏa dụng là hàm hạ tuyến tính (sublinear) về tài nguyên. Vấn đề trung tâm ở đây là sự to lớn của nguồn tài nguyên tiềm năng. Giả sử vũ trụ quan sát được thực sự là không có người ở, nó chứa nhiều hơn một thiên hà cho mỗi người còn sống. Hầu hết mọi người đều muốn được sử dụng một lượng tài nguyên tương đương một thiên hà chứ không phải một tờ vé số với cơ hội một phần tỷ trong việc trúng thưởng một tỷ thiên hà.³⁹ Với quy mô lớn của tài nguyên vũ trụ mà con người sở hữu, xem ra mọi người đều muốn một sự đảm bảo về việc ai cũng có phần, ngay cả

khi mỗi phần chỉ chiếm một tỷ lệ rất nhỏ. Điều quan trọng khi một vận hội như vậy sắp tới là không để mình bị loại ra khỏi cuộc chơi.

Lập luận dựa vào sự to lớn của “chiếc bánh” tài nguyên giả định rằng các ưu tiên là có thể thỏa mãn về mặt tài nguyên.⁴⁰ Giả định đó không nhất thiết đúng. Ví dụ, một số lý thuyết đạo đức nổi bật, bao gồm các lý thuyết hệ quả luận tập hợp đặc biệt, là tương ứng với các hàm thỏa dụng trung tính với rủi ro và tuyến tính về tài nguyên. Một tỷ thiên hà có thể được sử dụng để tạo ra cuộc sống hạnh phúc gấp tỷ lần so với một thiên hà. Vì thế, với một người thực dụng, chúng đáng giá hơn cả tỷ lần.⁴¹ Tuy vậy, về mặt tài nguyên, những mong muốn ích kỷ của con người bình thường là có thể thỏa mãn.

Tuyên bố cuối cùng này phải được bổ sung bởi hai nhận định quan trọng. Thứ nhất, nhiều người quan tâm đến thứ hạng. Nếu mỗi người đều muốn đứng đầu danh sách các cá nhân giàu có của Forbes, sẽ không có lượng tài nguyên nào đủ lớn để mang đến cho tất cả sự hài lòng trọn vẹn.

Nhận định thứ hai là nền tảng công nghệ hậu chuyển đổi sẽ cho phép biến tài nguyên vật chất thành một loạt sản phẩm chưa từng có, bao gồm một số loại hàng hóa hiện không có sẵn ở bất kỳ mức giá nào tuy được nhiều người đánh giá cao. Một tỷ phú không sống lâu hơn một ngàn lần so với một triệu phú. Tuy nhiên, trong thời đại của những bộ óc kỹ thuật số, tỷ phú có thể sở hữu công suất tính toán gấp ngàn lần, và do đó có thể tận hưởng tuổi thọ chủ quan lâu hơn gấp ngàn lần. Năng lực tinh thần, tương tự, cũng có thể được bán. Trong hoàn cảnh như vậy, với lượng vốn kinh tế có thể chuyển đổi thành hàng hóa thiết yếu với tốc độ không đổi ngay cả đối với mức độ của cải lớn hơn, lòng tham không bị ràng buộc sẽ có nhiều ý

nghĩa hơn so với thế giới ngày nay, nơi những người giàu (mà trong số đó có những người thiếu một trái tim bác ái) buộc phải tiêu pha vào máy bay, thuyền, các bộ sưu tập nghệ thuật hoặc hàng loạt dinh thự.

Liệu điều này có đồng nghĩa một người vị kỷ nên bỏ qua rủi ro liên quan đến tài nguyên hậu chuyển đổi của mình? Không hẳn vậy. Tài nguyên vật lý có thể không được chuyển đổi thành tuổi thọ hoặc hiệu năng tinh thần ở quy mô tùy ý. Nếu một cuộc sống phải diễn ra theo tuần tự, để người quan sát ở từng khoảnh khắc có thể nhớ các sự kiện trước đó và bị ảnh hưởng bởi các lựa chọn trước đó, thì cuộc sống của tâm trí kỹ thuật số có thể được mở rộng tùy ý mà không cần sử dụng một số lượng phép tính tuần tự ngày càng tăng. Nhưng các yếu tố vật lý lại giới hạn mức độ mà các tài nguyên có thể được chuyển đổi thành các phép tính tuần tự.⁴² Các giới hạn về tính toán tuần tự cũng có thể hạn chế một số khía cạnh của hiệu năng nhận thức trong việc mở rộng triệt để theo hàm hạ tuyến tính ra khỏi một nguồn tài nguyên tương đối khiêm tốn. Hơn nữa, chúng ta không thấy rõ việc một người vị kỷ sẽ (hoặc nên) bỏ qua rủi ro ngay cả đối với các số liệu kết quả có liên quan mang tính tiêu chuẩn cao như tuổi thọ chủ quan khi cuộc sống đã được điều chỉnh về chất lượng. Nếu đưa ra sự lựa chọn giữa việc chắc chắn kéo dài tuổi thọ thêm 2.000 năm và 1/10 khả năng kéo dài tuổi thọ thêm 30.000 năm, tôi nghĩ rằng hầu hết mọi người sẽ chọn phương án thứ nhất (ngay cả khi mỗi năm tuổi thọ đều có chất lượng như nhau).⁴³

Trong thực tế, lập luận cân nhắc lợi-hại ủng hộ việc phân phối lợi nhuận rộng rãi có lẽ là tương đối với chủ thể và phụ thuộc vào tình huống. Tuy nhiên, về tổng thể, mọi người sẽ có nhiều khả năng nhận

được (gần như tất cả) những gì họ muốn nếu tìm được cách thức tiến hành phân phối rộng rãi, và điều này luôn đúng kể cả trước khi tính đến việc cam kết phân phối rộng hơn sẽ có xu hướng thúc đẩy sự hợp tác, và do đó làm tăng khả năng tránh được thảm họa diệt vong. Do đó, việc ủng hộ phân phối rộng rãi dường như không chỉ là bắt buộc từ lập luận đạo đức mà cũng là điều nên làm từ lập luận cân nhắc lợi-hại.

Ngoài ra, chúng ta có một tập hợp các hệ quả khác của sự hợp tác cần đưa vào cân nhắc: khả năng hợp tác tiền chuyển đổi ảnh hưởng đến mức độ hợp tác hậu chuyển đổi. Giả sử nhân loại giải quyết được vấn đề kiểm soát. (Nếu vấn đề kiểm soát không được giải quyết, mức độ hợp tác hậu chuyển đổi sẽ không có nhiều ý nghĩa.) Có hai trường hợp cần xem xét. Đầu tiên là sự bùng nổ trí tuệ không tạo ra một động lực “được ăn cả” (có lẽ vì quá trình cất cánh xảy ra tương đối chậm). Trong trường hợp này, có thể nếu sự hợp tác tiền chuyển đổi có bất kỳ ảnh hưởng hệ thống nào đối với sự hợp tác hậu chuyển đổi thì ảnh hưởng đó sẽ là tích cực, có xu hướng thúc đẩy sự hợp tác tiếp theo. Các mối quan hệ hợp tác ban đầu có thể tồn tại và tiếp tục vượt qua quá trình chuyển đổi; ngoài ra, sự hợp tác tiền chuyển đổi có thể mang lại nhiều cơ hội hơn cho mọi người trong việc điều khiển sự phát triển theo hướng mong muốn (và có lẽ là hợp tác hơn) sau chuyển đổi.

Trường hợp thứ hai là bản chất của sự bùng nổ trí tuệ có thúc đẩy một động lực “được ăn cả” (có lẽ vì quá trình cất cánh là tương đối nhanh). Trong trường hợp này, nếu không có sự hợp tác sâu rộng trước khi cất cánh, một thể đơn nhất có khả năng sẽ xuất hiện (một dự án duy nhất sẽ trải qua quá trình chuyển đổi, và đến một lúc

nào đó sẽ có được ưu thế chiến lược quyết định cùng với siêu trí tuệ). Theo định nghĩa, một thể đơn nhất là một trật tự xã hội có tính hợp tác cao.⁴⁴

Việc không có sự hợp tác sâu rộng tiền chuyển đổi sẽ dẫn đến một mức độ hợp tác hậu chuyển đổi cực lớn. Ngược lại, một mức độ hợp tác cao hơn một chút trước khi xảy ra sự bùng nổ trí tuệ sẽ mở ra nhiều kết quả khả dĩ hơn. Các dự án hợp tác có thể đồng bộ hóa tiến bộ của mình để đảm bảo một sự chuyển đổi đồng thời mà không có bất kỳ dự án nào có được ưu thế chiến lược quyết định. Hoặc các nhóm tài trợ khác nhau có thể hợp nhất các nỗ lực của họ vào một dự án duy nhất, trong khi không cho phép dự án đó hình thành thể đơn nhất. Ví dụ, ta có thể tưởng tượng nhiều quốc gia cùng xây dựng một dự án khoa học chung để phát triển siêu trí tuệ máy, nhưng không cho phép dự án này phát triển thành bất cứ thứ gì như Liên Hợp Quốc siêu cấp, mà thay vào đó sẽ lựa chọn việc duy trì trật tự thế giới có tính bè phái đã tồn tại trước đó.

Do đó, đặc biệt trong trường hợp cạnh tranh nhanh, có khả năng sự hợp tác tiền chuyển đổi lớn hơn sẽ dẫn đến sự hợp tác hậu chuyển đổi mức độ thấp hơn. Tuy nhiên, ở mức độ mà các thực thể hợp tác có thể định hình kết quả, họ chỉ có thể cho phép sự không hợp tác xuất hiện hoặc tiếp diễn nếu thấy trước rằng không có hậu quả thảm khốc nào xảy ra từ tình trạng bè phái hậu chuyển đổi. Do đó, các kịch bản mà trong đó, hợp tác tiền chuyển đổi dẫn đến giảm mức độ hợp tác hậu chuyển đổi có thể chủ yếu là các kịch bản khi giảm hợp tác hậu chuyển đổi là vô hại.

Nói chung, một mức độ hợp tác hậu chuyển đổi lớn hơn xem ra là đáng mong đợi. Nó sẽ làm giảm nguy cơ xuất hiện một động lực sai

lệch, trong đó cạnh tranh kinh tế và dân số mở rộng nhanh chóng sẽ dẫn đến một tình trạng kiểu “bẫy Malthus”, trong đó chọn lọc tiến hóa làm xói mòn các giá trị của con người và lựa chọn các hình thức phi đạo đức, hoặc trong đó các cường quốc đối thủ phải chịu những thất bại về phối hợp khác như chiến tranh và các cuộc đua công nghệ. Vấn đề cuối cùng, triển vọng của các cuộc đua công nghệ, có thể đặc biệt rắc rối nếu quá trình chuyển đổi hướng tới một dạng trung gian của trí tuệ máy (giả lập hoàn chỉnh não bộ), vì nó sẽ tạo ra một động lực chạy đua mới có ảnh hưởng xấu đến cơ hội giải quyết vấn đề kiểm soát cho quá trình chuyển đổi thứ hai sang một dạng trí tuệ máy tiên tiến hơn (trí tuệ nhân tạo).

Ở trên, chúng ta đã bàn về cách mà sự hợp tác có thể làm giảm xung đột trong khi xảy ra sự bùng nổ trí tuệ, tăng khả năng giải quyết vấn đề kiểm soát và cải thiện cả tính chính đáng về đạo đức cùng mức độ mong muốn về lợi-hại của việc phân bổ nguồn lực. Vì vậy, những lợi ích này của sự hợp tác có thể được bổ sung một luận điểm: sự hợp tác tiền chuyển đổi sâu rộng hơn có thể giúp giải quyết các vấn đề phối hợp quan trọng trong thời kỳ hậu chuyển đổi.

Cùng nhau làm việc

Có nhiều hình thức hợp tác khác nhau, tùy thuộc vào quy mô của các thực thể hợp tác. Ở quy mô nhỏ, các nhóm AI đơn lẻ tin rằng mình sẽ cạnh tranh với nhau có thể lựa chọn việc chung sức nỗ lực.⁴⁵ Các tập đoàn có thể hợp nhất hoặc đầu tư chéo. Ở quy mô lớn hơn, các quốc gia có thể tham gia vào một dự án quốc tế lớn. Đã có tiền lệ cho sự hợp tác quốc tế quy mô lớn về khoa học và công nghệ (như CERN, Dự án bộ gen người và Trạm vũ trụ quốc tế), nhưng một dự án quốc tế phát triển siêu trí tuệ an toàn sẽ đặt ra một thách thức

khác vì những vấn đề về an ninh. Nó sẽ phải được cấu thành không phải như một sự hợp tác học thuật mở mà là một sự liên kết hành động được kiểm soát chặt chẽ. Có lẽ các nhà khoa học liên quan sẽ phải cách ly và không được giao tiếp với bên ngoài trong suốt thời gian dự án, ngoại trừ thông qua một kênh liên lạc được xem xét cẩn thận. Mức độ bảo mật cần thiết là gần như không thể đạt được ở thời điểm hiện tại, nhưng những tiến bộ trong công nghệ phát hiện, giám sát có thể làm cho việc đó trở nên khả thi vào cuối thế kỷ này. Ta cũng cần ghi nhớ rằng sự hợp tác rộng rãi không nhất thiết là một số lượng lớn các nhà nghiên cứu sẽ tham gia vào dự án, mà chỉ đơn giản là nhiều người sẽ có tiếng nói trong dự án. Về nguyên tắc, một dự án có thể có sự hợp tác sâu rộng tối đa, với toàn bộ nhân loại đều tham gia tài trợ (và có thể do Đại hội đồng Liên Hợp Quốc đứng ra đại diện), nhưng chỉ cần một nhà khoa học duy nhất để thực hiện công việc.⁴⁶

Có một lý do để bắt đầu hợp tác càng sớm càng tốt, đó là tận dụng việc chúng ta bị một bức màn vô thức che đi các thông tin cụ thể về việc dự án nào sẽ có siêu trí tuệ trước. Càng gần đến đích, chúng ta sẽ càng chắc chắn hơn về cơ hội tương đối của các dự án cạnh tranh; và có thể sẽ càng khó hơn để thuyết phục người dẫn đầu tham gia một dự án hợp tác phân phối lợi ích cho toàn nhân loại. Mặt khác, khó có thể thiết lập sự hợp tác chính thức trên phạm vi toàn thế giới trước khi triển vọng siêu trí tuệ được công nhận rộng rãi hơn nhiều so với hiện tại, và trước khi có một con đường rõ ràng dẫn đến việc tạo ra siêu trí tuệ máy. Hơn nữa, ở mức độ mà sự hợp tác sẽ thúc đẩy tiến bộ trên con đường đó, như chúng ta đã thảo luận, nó thực sự có thể phản tác dụng về mặt an toàn.

Do đó, hình thức hợp tác lý tưởng cho hiện tại có thể là một hình thức mà ban đầu không đòi hỏi các thỏa thuận chính thức cụ thể và không xúc tiến những tiến bộ trong trí tuệ máy. Một đề xuất phù hợp với các tiêu chí này là chúng ta cần đưa ra một chuẩn mực đạo đức phù hợp, thể hiện cam kết của mình đối với ý tưởng rằng siêu trí tuệ nên vì lợi ích chung. Một định mức như vậy có thể được xây dựng như sau:

Nguyên tắc lợi ích chung:

Siêu trí tuệ chỉ nên được phát triển vì lợi ích của toàn nhân loại và phục vụ các lý tưởng đạo đức phổ quát.⁴⁷

Việc quy định từ giai đoạn đầu rằng tiềm năng to lớn của siêu trí tuệ thuộc về toàn nhân loại sẽ mang lại thêm thời gian để củng cố chuẩn mực này.

Nguyên tắc lợi ích chung không loại trừ các ưu đãi thương mại cho các cá nhân hoặc doanh nghiệp hoạt động trong các lĩnh vực liên quan. Ví dụ, một công ty có thể đáp ứng lời kêu gọi chia sẻ toàn cầu về lợi ích của siêu trí tuệ bằng cách áp dụng một “điều khoản lợi nhuận bất ngờ” (windfall clause) với hiệu lực là toàn bộ lợi nhuận với một mức trần rất cao (ví dụ, một nghìn tỷ đô la hàng năm) sẽ được phân phối theo cách thông thường cho các cổ đông của công ty và các bên yêu cầu pháp lý khác, và những lợi nhuận vượt quá mức trần đó sẽ được phân phối đều cho toàn nhân loại (hoặc theo cách khác, dựa trên các tiêu chí đạo đức phổ quát). Việc áp dụng điều khoản này sẽ rất tốn kém, bất kỳ công ty nào cũng cực kỳ khó vượt quá được một mức trần lợi nhuận ngất ngưỡng như vậy (và các kịch bản có xác suất thấp như vậy thường không có vai trò gì trong quyết định của các nhà quản lý và nhà đầu tư của công ty). Tuy nhiên, việc

áp dụng rộng rãi sẽ mang lại cho nhân loại một sự đảm bảo có giá trị (trong chừng mực mà các cam kết có thể được tin cậy) rằng nếu có một doanh nghiệp tư nhân nào đó “trúng số độc đắc” với sự bùng nổ trí tuệ, mọi người sẽ được chia sẻ hầu hết các lợi ích. Ý tưởng tương tự có thể được áp dụng cho các thực thể khác ngoài các công ty. Ví dụ, các quốc gia có thể đồng ý rằng nếu GDP của bất kỳ quốc gia nào vượt qua một tỷ trọng lớn (giả sử, 90%) của GDP thế giới, thì phần dư này nên được phân bổ đều cho tất cả.⁴⁸

Nguyên tắc lợi ích chung (và các biện pháp thực hiện cụ thể, chẳng hạn như điều khoản lợi nhuận bất ngờ) có thể được chấp nhận ban đầu như một cam kết đạo đức tự nguyện của các cá nhân và tổ chức có trách nhiệm hoạt động trong các lĩnh vực liên quan đến trí tuệ máy. Sau đó, nó có thể được tán thành bởi một tập hợp rộng hơn các thực thể và được ban hành thành luật và hiệp ước. Một “công thức” mơ hồ, chẳng hạn như công thức được đưa ra ở đây, có thể được áp dụng như một điểm khởi đầu; nhưng cuối cùng nó sẽ cần được mài giũa thành một tập hợp các yêu cầu có thể kiểm chứng cụ thể.

Chương 15

Giai đoạn nước rút

Chúng ta nhận ra mình đang chìm trong một sự phức tạp về chiến lược và bị bao quanh bởi một màn sương dày đặc của sự không chắc chắn. Mặc dù nhiều cân nhắc đã được phân biệt, các chi tiết và mối quan hệ của chúng vẫn rất mơ hồ, và có thể có những yếu tố khác mà chúng ta chưa từng nghĩ đến. Chúng ta phải làm gì trong tình cảnh khó khăn này?

Triết học có “deadline”

Một đồng nghiệp của tôi hay nói rằng giải thưởng Field (danh hiệu cao nhất trong toán học) chỉ ra hai điều về người nhận: anh ta có khả năng hoàn thành một việc quan trọng, và anh ta đã không làm thế. Tuy khắc nghiệt nhưng đây là một sự thật.

Hãy nghĩ về một “khám phá” như một hành động khiến cho thông tin xuất hiện ở một thời điểm sớm hơn. Giá trị của khám phá không tương đương giá trị của thông tin được khám phá, mà là giá trị của thời điểm mà ta có được thông tin đó. Một nhà khoa học hoặc một nhà toán học có thể thể hiện kỹ năng tuyệt vời khi là người đầu tiên tìm ra một giải pháp mà nhiều người không thể; tuy nhiên, nếu vấn đề đang nào cũng sẽ được giải quyết thì công trình đó có lẽ không mang lại nhiều lợi ích cho thế giới. Có những trường hợp mà trong đó, giải pháp đến sớm hơn một chút là vô cùng có giá trị, nhưng sẽ là

hợp lý nhất nếu giải pháp được đưa vào áp dụng ngay lập tức (được triển khai trong thực tiễn hoặc làm nền tảng để nghiên cứu lý thuyết sâu hơn).

Trong trường hợp thứ hai, khi một giải pháp chỉ được sử dụng ngay lập tức với vai trò nền tảng lý thuyết, nó sẽ chỉ có giá trị rất lớn nếu xuất hiện sớm hơn khi các công trình nghiên cứu mà nó cho phép thực hiện vừa quan trọng vừa cấp bách.¹

Vì thế, câu hỏi không phải là liệu kết quả được phát hiện bởi người nhận giải thưởng Field có quan trọng hay không (bất kể mang tính công cụ hay vì bản thân kiến thức đó). Thay vào đó, câu hỏi là có quan trọng không nếu người nhận giải cho phép công bố kết quả vào một ngày trước đó. Giá trị của việc dịch chuyển thời điểm này nên được so sánh với giá trị mà một bộ óc toán học tầm cỡ thế giới có thể tạo ra bằng cách làm việc trên một thứ khác. Ít nhất là trong một số trường hợp, giải thưởng Field có thể phản ánh việc dành cả đời để giải quyết nhằm vấn đề, chẳng hạn như một vấn đề chỉ hấp dẫn vì độ khó.

Những lời chua chát này cũng có thể được hướng vào các lĩnh vực khác, chẳng hạn như triết học hàn lâm. Triết học nói về một số vấn đề có liên quan đến giảm thiểu rủi ro diệt vong, thứ mà chúng ta đã bàn đến vài lần. Tuy nhiên, cũng có những lĩnh vực nhỏ trong triết học không có mối liên hệ rõ ràng với rủi ro diệt vong hoặc bất kỳ mối quan tâm thực tế nào. Như với toán học thuần túy, một số vấn đề mà nghiên cứu triết học có thể coi là quan trọng về mặt nội tại, nghĩa là con người có lý do để quan tâm đến chúng một cách độc lập với bất kỳ ứng dụng thực tế nào; chẳng hạn, chúng ta muốn tìm hiểu bản chất của thực tại vì chính nó. Thế giới có thể sẽ kém huy hoàng hơn

nếu không ai nghiên cứu siêu hình học, vũ trụ học hay lý thuyết dây. Tuy nhiên, viễn cảnh khởi đầu của một sự bùng nổ trí tuệ đã chiếu một luồng sáng mới đến công cuộc tìm kiếm trí tuệ cổ xưa này.

Triển vọng bây giờ cho thấy rằng tiến bộ triết học có thể được tối đa hóa bằng một con đường gián tiếp chứ không cần ngay lập tức đưa ra triết lý. Một trong nhiều nhiệm vụ mà siêu trí tuệ (hoặc thậm chí chỉ là trí tuệ con người được tăng cường vừa phải) sẽ vượt trội hơn các nhà tư tưởng hiện nay là trả lời các câu hỏi cơ bản trong khoa học và triết học. Sự phản ánh này cho thấy một chiến lược “kìm hãm sự sung sướng”. Chúng ta có thể hoãn trả lời một số câu hỏi trong một thời gian ngắn và giao nhiệm vụ đó cho những người kế nhiệm có năng lực cao hơn để tập trung vào một thách thức cấp bách hơn: gia tăng cơ hội thực sự có được những “người kế nhiệm” có năng lực đó. Đây chính là triết học và toán học với hiệu quả cao.²

Những việc cần làm

Do đó, chúng ta muốn tập trung vào các vấn đề không chỉ quan trọng mà còn cấp bách theo nghĩa các giải pháp cho chúng là cần thiết trước khi xảy ra sự bùng nổ trí tuệ. Chúng ta cũng cần lưu ý không làm việc với các vấn đề có “giá trị âm” (nghĩa là giải quyết chúng là có hại). Chẳng hạn, một số vấn đề kỹ thuật trong lĩnh vực trí tuệ nhân tạo có thể có giá trị âm vì giải quyết chúng sẽ tăng tốc độ phát triển trí tuệ máy mà không làm gì nhiều để thúc đẩy sự phát triển của các phương pháp kiểm soát có thể giúp con người sống sót và có được lợi ích từ cuộc cách mạng trí tuệ máy.

Xác định các vấn đề cấp bách, quan trọng và có thể hoàn toàn được coi là có “giá trị dương” quả không đơn giản. Sự không chắc chắn về chiến lược xung quanh việc giảm thiểu rủi ro diệt vong đồng

nghĩa chúng ta phải lo lắng rằng ngay cả những can thiệp có chủ ý tốt cũng có thể không chỉ không hiệu quả mà còn phản tác dụng. Để hạn chế nguy cơ làm điều có hại hoặc sai trái về đạo đức, chúng ta nên ưu tiên xử lý các vấn đề có giá trị dương mạnh mẽ (nghĩa là giải pháp cho chúng sẽ đóng góp tích cực trong một loạt các kịch bản) và sử dụng các biện pháp hoàn toàn chính đáng (nghĩa là, chấp nhận được từ một loạt các quan điểm đạo đức).

Chúng ta có một ước nguyện khác cần cân nhắc khi lựa chọn các vấn đề ưu tiên. Chúng ta muốn làm việc trên các vấn đề đáp ứng tốt với những nỗ lực của mình trong việc giải quyết chúng. Các vấn đề đáp ứng tốt là các vấn đề có thể được giải quyết nhanh hơn nhiều, hoặc giải quyết ở mức độ lớn hơn nhiều nếu nỗ lực lớn hơn đôi chút. Việc khích lệ lòng tốt trên thế giới là một vấn đề quan trọng và cấp bách, một vấn đề mà hơn nữa, còn có giá trị dương mạnh mẽ nhưng chưa có ý tưởng đột phá về cách thực hiện. (Có lẽ là đây vấn đề có mức độ đáp ứng khá thấp.) Đạt được hòa bình thế giới, tương tự, là việc rất được trông đợi; nhưng khi xem xét rất nhiều nỗ lực đã nhắm vào nó và những trở ngại lớn cho sự ra đời của một giải pháp nhanh chóng, xem ra những đóng góp của một số ít cá nhân sẽ khó mang lại sự khác biệt đáng kể.

Để giảm thiểu rủi ro của cuộc cách mạng trí tuệ máy, chúng tôi sẽ đề xuất hai mục tiêu có vẻ đáp ứng tốt nhất cho tất cả những ước nguyện đó: phân tích chiến lược và nâng cao năng lực. Chúng ta có thể tương đối tự tin về “dấu” của các tham số này: hiểu biết sâu sắc hơn về chiến lược và năng lực cao hơn sẽ tốt hơn. Hơn nữa, các tham số này cũng có mức độ đáp ứng cao: một khoản đầu tư gia tăng nhỏ có thể tạo ra sự khác biệt tương đối lớn. Có được hiểu biết

sâu sắc và năng lực cũng là việc cấp bách bởi vì việc thúc đẩy sớm các tham số này có thể kết hợp, làm cho những nỗ lực tiếp theo trở nên hiệu quả hơn. Ngoài hai mục tiêu rộng lớn này, chúng tôi sẽ chỉ ra một số mục tiêu tiềm năng khác đáng để đầu tư tìm kiếm sáng kiến.

Đi tìm ánh sáng chiến lược

Trong bối cảnh phức tạp và không chắc chắn, phân tích là một việc làm có giá trị kỳ vọng đặc biệt cao.³ Tình hình chiến lược được soi sáng sẽ giúp chúng ta đặt mục tiêu hiệu quả hơn cho các biện pháp can thiệp tiếp theo. Phân tích chiến lược là đặc biệt cần thiết khi chúng ta hoàn toàn không chắc chắn không chỉ về một số chi tiết của một số vấn đề phụ mà còn về phẩm chất chính yếu của các vấn đề trung tâm. Đối với nhiều tham số chính, chúng ta hoàn toàn không chắc chắn về dấu của chúng, nghĩa là, chúng ta không biết hướng thay đổi nào là đáng trông đợi và hướng nào không. Sự thiếu hiểu biết của chúng ta có thể không bù đắp được. Lĩnh vực này chưa được khai thác nhiều, và những ý tưởng chiến lược sâu sắc cận kề có thể vẫn đang chờ đợi được khám phá.

Ở đây, chúng ta nói về “phân tích chiến lược” như là một cuộc tìm kiếm những cân nhắc quan trọng: những ý tưởng hoặc lập luận có khả năng thay đổi quan điểm của chúng ta không chỉ về cấu trúc chi tiết của quá trình thực thi mà còn về một quang cảnh chung của mức độ đáng trông đợi.⁴ Một cân nhắc quan trọng bị bỏ lỡ có thể làm hỏng những nỗ lực quả cảm nhất của chúng ta và biến chúng thành có hại, như một người lính chọn sai phe vậy. Việc tìm kiếm những cân nhắc quan trọng (phải tìm ra cả các vấn đề tiêu chuẩn và mô tả) thường sẽ đòi hỏi phải vượt qua ranh giới giữa các ngành và các

lĩnh vực kiến thức khác nhau. Vì loại hình nghiên cứu này không có phương pháp chính thống nên tư duy ban đầu nhất thiết sẽ khó khăn.

Nâng cao năng lực

Một hoạt động có giá trị cao khác, với sự mạnh mẽ không kém phân tích chiến lược trong việc mang lại lợi ích trong hàng loạt kịch bản khác nhau, là phát triển một nền tảng hỗ trợ vững chắc và có tầm nhìn xa. Một nền tảng như vậy có thể lập tức cung cấp các nguồn lực cho nghiên cứu và phân tích. Nếu và khi các ưu tiên khác trở nên rõ ràng, tài nguyên có thể được chuyển hướng tương ứng. Vì thế, một nền tảng hỗ trợ là một năng lực phổ quát có thể được định hướng áp dụng bởi các kiến thức mới khi chúng xuất hiện.

Một tài sản có giá trị sẽ là một mạng lưới các nhà tài trợ bao gồm các cá nhân sẵn sàng tham gia hoạt động từ thiện hợp lý, biết về rủi ro diệt vong và nhận thức rõ về các biện pháp giảm thiểu. Điều đặc biệt mong muốn là các nhà tài trợ ban đầu này phải sắc sảo và vị tha, vì họ có cơ hội định hình văn hóa của lĩnh vực này trước khi những mối quan tâm vụ lợi thông thường xuất hiện và củng cố vị trí của mình. Do đó, trọng tâm của những bước đi đầu tiên này nên nhắm đến việc chiêu mộ đúng người. Từ bỏ một số tiến bộ công nghệ ngắn hạn để chiêu mộ các cá nhân toàn tâm toàn ý với vấn đề an toàn và có định hướng đi tìm sự thật (và có khả năng thu hút những cá nhân khác giống thế).

Một biến số quan trọng là chất lượng của “nhận thức luận xã hội” của lĩnh vực AI và các dự án hàng đầu trong đó. Khám phá những cân nhắc quan trọng là có giá trị, nhưng chỉ khi nó ảnh hưởng đến hành động. Điều này không thể luôn luôn được coi là đương nhiên.

Hãy tưởng tượng một dự án đầu tư hàng triệu đô la và nhiều năm để phát triển một AI nguyên mẫu và sau khi vượt qua nhiều thách thức kỹ thuật, hệ thống cuối cùng đã bắt đầu cho thấy sự tiến bộ thực sự. Có một khả năng AI đó sẽ trở nên hữu dụng và mang lại lợi nhuận mà chỉ cần tiến hành thêm đôi chút nghiên cứu. Bây giờ một cân nhắc quan trọng được phát hiện, chỉ ra rằng một cách tiếp cận hoàn toàn khác sẽ có đôi chút an toàn hơn. Liệu dự án đó có “mổ bụng tự sát” như võ sĩ đạo, từ bỏ thiết kế không an toàn và tất cả những tiến bộ đã có? Hay dự án sẽ hành xử như một con bạch tuộc đang sợ hãi, phun ra một đám “mực” của sự hoài nghi với hy vọng trốn tránh cuộc tấn công? Một dự án lựa chọn hành xử như võ sĩ đạo trong tình huống như vậy sẽ là một dự án được ưa thích.⁵ Tuy nhiên, việc xây dựng các quy trình và thể chế sẵn sàng thực hiện việc “mổ bụng tự sát” này dựa trên những lập luận không chắc chắn và tính suy đoán là không dễ dàng. Một khía cạnh khác của nhận thức luận xã hội là việc quản lý thông tin nhạy cảm, cụ thể là khả năng tránh rò rỉ những thông tin cần được giữ bí mật. (Việc giữ kín thông tin có thể đặc biệt khó khăn đối với các nhà nghiên cứu hàn lâm, đã quen với việc liên tục công bố kết quả trên mọi phương tiện sẵn có.)

Các biện pháp cụ thể

Ngoài các mục tiêu chung về chiến lược và năng lực, một số mục tiêu cụ thể hơn cũng có thể mang lại cơ hội hành động tiết kiệm chi phí. Một trong số đó là tiến bộ trong việc xử lý những thách thức kỹ thuật về an toàn của trí tuệ máy. Khi theo đuổi mục tiêu này, cần thận trọng trong việc xử lý các mối nguy hại về thông tin. Một số công việc hữu ích trong giải quyết vấn đề kiểm soát cũng sẽ hữu ích trong giải

quyết vấn đề năng lực. Các công việc khiến sự bùng nổ của AI được châm ngòi sớm hơn có thể dễ dàng trở nên tiêu cực.

Một mục tiêu cụ thể khác là thúc đẩy áp dụng các “thông lệ tốt nhất” (best practice) trong đội ngũ nghiên cứu AI. Bất cứ tiến bộ nào đã được thực hiện đối với vấn đề kiểm soát cần phải được phổ biến. Một số hình thức thử nghiệm tính toán, đặc biệt nếu liên quan đến quá trình tự cải thiện đệ quy mạnh mẽ, cũng có thể đòi hỏi áp dụng việc kiểm soát khả năng để giảm thiểu rủi ro xảy ra một quá trình cất cánh bất ngờ. Mặc dù việc triển khai thực tế các phương pháp an toàn hiện nay không hẳn là phù hợp, nhưng nó sẽ ngày càng trở nên phù hợp khi công nghệ phát triển. Không còn quá sớm để kêu gọi giới chuyên môn đưa ra một cam kết về an toàn, bao gồm cả việc tán thành nguyên tắc lợi ích chung và lời hứa về việc tăng cường an toàn nếu và khi viễn cảnh siêu trí tuệ máy bắt đầu trở nên hiển hiện. Những lời nói sáo rỗng là không đủ và sẽ không thể khiến cho một công nghệ nguy hiểm trở nên an toàn: nhưng lời nói có thể sẽ đi đôi với tư duy.

Các cơ hội khác đôi khi cũng có thể xuất hiện để thúc đẩy một tham số quan trọng nào đó, chẳng hạn như giảm thiểu một số rủi ro diệt vong khác, thúc đẩy quá trình tăng cường nhận thức sinh học và cải thiện trí tuệ tập thể, hoặc thậm chí khiến cho tình hình chính trị thế giới trở nên hài hòa hơn.

Hãy khơi dậy những bản chất tốt đẹp nhất

Trước viễn cảnh bùng nổ trí tuệ, con người chúng ta giống như những đứa trẻ đang chơi với một quả bom. Đó là sự không phù hợp giữa sức mạnh của thứ “đồ chơi” đó và sự non nớt trong hành vi của con người. Siêu trí tuệ là một thách thức mà hiện tại chúng ta chưa

sẵn sàng và sẽ không sẵn sàng đối phó trong một thời gian dài. Chúng ta hầu như không có thông tin về thời điểm phát nổ, mặc dù tiếng đồng hồ đếm ngược vẫn vang vọng bên tai.

Với một đứa trẻ cầm quả bom trong tay, điều hợp lý cần làm là đặt nó xuống một cách nhẹ nhàng, nhanh chóng ra khỏi khu vực nguy hiểm và liên lạc với người lớn ở nơi gần nhất. Tuy nhiên, ở đây chúng ta có rất nhiều đứa trẻ, và mỗi đứa trẻ đều có quyền truy cập vào một cơ chế kích hoạt độc lập. Cơ hội mà ở đó tất cả chúng ta đều thấy cần phải đặt thứ đó xuống gần như là con số không. Một vài tên nhóc ngốc nghếch luôn muốn bấm nút kích nổ chỉ để xem điều gì sẽ đến.

Chúng ta cũng không thể có được sự an toàn bằng cách chạy trốn, vì sự bùng nổ trí tuệ sẽ thổi bay mọi thứ. Chúng ta cũng không nhìn thấy “người lớn” nào xung quanh cả.

Trong tình huống này, toàn bộ cảm giác phấn khích sẽ không còn nữa mà thay vào đó sẽ là hoảng loạn và sợ hãi; nhưng thái độ phù hợp nhất sẽ là quyết định đầu tư vào năng lực càng nhiều càng tốt, giống như khi chuẩn bị cho một kỳ thi khó khăn sẽ chấp cánh hoặc huỷ hoại ước mơ.

Thái độ này không phải là một kiểu “cuồng tín”. Sự bùng nổ trí tuệ có thể cách chúng ta nhiều thập kỷ trong tương lai. Hơn nữa, thách thức mà chúng ta phải đối mặt một phần là giữ lấy nhân tính của mình: duy trì được sự vững chắc, lương tri và tính đúng đắn ngay cả khi vấn đề phi tự nhiên và phi nhân tính nhất này xảy đến. Chúng ta cần phải vận dụng tất cả nguồn lực của con người để giải quyết vấn đề này.

Nhưng, chúng ta cũng không nên đánh mất những gì có ý nghĩa toàn cầu. Thông qua những chuyện vặt vãnh hằng ngày, chúng ta có thể nhận thức, suy mơ hồ, về nhiệm vụ thiết yếu của thời đại. Trong cuốn sách này, chúng tôi đã cố gắng phân biệt thêm một vài nét chính của một viễn cảnh vẫn còn tương đối mơ hồ và được định nghĩa một cách tiêu cực – một viễn cảnh đặt ra ưu tiên hàng đầu về đạo đức (ít nhất là từ quan điểm phi cá nhân và theo chủ nghĩa thực) cho việc giảm thiểu rủi ro diệt vong và duy trì quỹ đạo nền văn minh dẫn đến việc tài nguyên vũ trụ của nhân loại được sử dụng một cách hồ hởi và đầy lòng trắc ẩn.

Lời kết

Kể từ khi xuất bản phiên bản bìa cứng của cuốn sách này, một số quan điểm đã thay đổi. Việc coi siêu trí tuệ là một chủ đề không ngở ngẩn nữa đã trở nên dễ dàng hơn – cụ thể là việc xem xét một cách nghiêm túc quan điểm cho rằng sự chuyển đổi trí tuệ máy có thể sẽ xảy ra ở thế kỷ này, rằng sự chuyển đổi ấy có thể coi là một trong những sự kiện quan trọng nhất trong lịch sử, rằng điều này có thể đi kèm với một số rủi ro diệt vong cũng như triển vọng phát triển to lớn, và rằng việc thực hiện trước một số công việc với mục đích xem xét liệu chúng ta có thể làm được gì để thu gọn phần dư thừa của kết quả có triển vọng cũng sẽ là một điều khôn ngoan. Cứ cho là như vậy thì vẫn còn đó hình ảnh thực tế của những người cuối cùng đang cười nhạo mỗi một nỗ lực tham gia vào chủ đề này trên mặt báo. Tuy nhiên, vượt ra khỏi sự không hòa hợp đại chúng này, nếu người ta “về” tai lên và nghe ngóng đúng hướng, họ sẽ có thể nghe thấy được tiếng thì thầm nho nhỏ của một cuộc đối thoại “trưởng thành” hơn.

Tiến bộ kỹ thuật trong học máy đã và đang diễn ra nhanh hơn so với dự kiến. Một loạt các ý tưởng cần khám phá đã được mở ra nhờ những phát triển gần đây: các máy Turing thần kinh, học tăng cường sâu, tối ưu hóa siêu tham số Bayes, lưới LSTMs, các mạng lưới bộ nhớ, bộ mã hóa tự động biến phân, phép nhúng vector cấp độ câu, mạng lưới đối kháng tạo sinh, mô hình tạo sinh dựa trên mối quan

tâm (attention-based), các đường hướng khác nhau trong lập trình xác suất – để chỉ ra một phần nhỏ trong số các nguồn của tin tức đương đại. Học sâu đã và đang là trọng tâm của phần lớn sự phấn khích. Nhờ có sự kết hợp của các máy tính nhanh hơn, các bộ dữ liệu lớn hơn và những sự tinh chỉnh thuật toán, các phương pháp học sâu mà về cơ bản là những mạng lưới thần kinh đa lớp đã bắt đầu tiệm cận (và trong một số trường hợp còn vượt quá) hiệu suất của con người trong nhiều tác vụ nhận thức, bao gồm nhận dạng chữ viết tay, nhận dạng hình ảnh và chú thích hình ảnh, nhận dạng giọng nói và nhận dạng khuôn mặt. Các phương pháp học sâu cũng đã đạt được những kết quả mạnh mẽ trong lĩnh vực dịch ngôn ngữ tự nhiên và trong một số vấn đề về phân tích dữ liệu khoa học. Khả năng ẩn bên dưới những màn trình diễn này, các thuật toán chung có thể học được các phép biểu diễn phân tán trừu tượng từ những dữ liệu cảm giác thô mà không cần tới các đặc tính thủ công hay tri thức đặc thù của con người, có thể trở thành một khối cơ bản quan trọng trong việc xây dựng nên các khả năng phức tạp hơn.

Rất nhiều khả năng hiện có đã đạt đến ngưỡng hữu ích. Điều này đồng nghĩa với việc chúng ta lập tức có được kết quả từ những tiến bộ xa hơn, khi chúng được trực tiếp chuyển đổi thành những cải tiến trong các sản phẩm có giá trị thương mại. Nếu bắt đầu với một hệ thống nhận dạng giọng nói có độ chính xác thấp đến mức không thể sử dụng và cải tiến nó một chút, thì hệ thống mà bạn có vẫn vô dụng. Nhưng, khi bắt đầu với một hệ thống đủ tốt để có thể được phát triển rộng rãi, thì sự cải thiện dù chỉ 1% cũng có thể đáng giá một tỷ đô la. Học máy đang mang lại một cảm giác ly kỳ với nhiều con đường để

khám phá – cảm giác này giờ đây ngày càng mạnh mẽ nhờ các động cơ thương mại. Điều đó đang thu hút thêm nhiều vốn và tài năng.

Việc làn sóng phấn khích hiện tại sẽ đi bao xa chỉ mang tính dự đoán. Cuốn sách này không hề lập luận rằng một sự bùng nổ trí tuệ sắp xảy ra, hoặc thế giới rộng lớn đã đánh giá thấp tốc độ tiến bộ. Rất có thể, những cải tiến có thể đạt được bằng cách nhân rộng và tinh chỉnh bất kỳ phương pháp tiếp cận hiện tại nào sẽ sớm trở nên bình ổn. Chắc chắn rằng, có thêm các ý tưởng đột phá sẽ là cần thiết để đi hết con đường (đó là các ý tưởng có thể hoặc không thể được đưa ra một cách kịp thời). Tuy nhiên, tôi sẽ ngạc nhiên đôi chút khi thấy sự tái xuất của một mùa đông AI khắc nghiệt như những gì mà lĩnh vực này đã trải qua trong quá khứ. Dường như có nhiều khả năng uy tín và các nguồn tài trợ liên tục sẽ xuất hiện từ thời điểm này, vì công cuộc nghiên cứu đã đạt thành tựu đủ để AI đứng vững như một ngành nghiên cứu lý thuyết đáng giá, và đồng thời cũng là một ngành kỹ thuật hữu ích. (Dự đoán này, tất nhiên, hoàn toàn phù hợp với sự bùng nổ của các bong bóng đầu tư, cũng như sự thất bại của hầu hết các dự án đơn lẻ trong việc đáp ứng kỳ vọng của người đề xướng.)

Cũng đã có một số tiến bộ trong nghiên cứu về cách cải thiện tỷ lệ của một kết quả thuận lợi. Về mặt lý thuyết, một số ý tưởng mới đã được giới thiệu (ví dụ: các “tác tử được định hướng theo phê duyệt” của Paul Christiano), và rất đáng để khám phá thêm. Về mặt tổ chức cũng vậy: tình hình đã được cải thiện đôi chút (sự cải thiện là rất lớn về mặt tương đối nhưng lại đang bắt đầu từ một mức cơ bản rất thấp). Việc thừa nhận vấn đề kiểm soát là một chủ đề phù hợp để điều tra đã lan rộng đến một phần lớn hơn của các cộng đồng nghiên

cứu có liên quan, và người ta cũng đã thừa nhận hơn rằng các tác động trong tương lai của trí tuệ máy xứng đáng có được một sự quan tâm có hệ thống hơn.

Mặc dù đáng chú ý vì đã diễn ra trong một khoảng thời gian ngắn, nhưng sự tiến bộ này trong việc củng cố các vấn đề an toàn và tác động của AI không nên được thổi phồng. Thật vậy, nhiều nguồn tài trợ đang chảy vào lĩnh vực này; nhưng nó vẫn còn ít hơn hai đến ba bậc so với việc đơn giản là làm cho máy móc thông minh hơn. Người ta đã hứng thú hơn khi nghĩ về các chuỗi tiến bộ trong trí tuệ máy; nhưng phần lớn trong số này cuối cùng lại tập trung vào các mối quan tâm ngắn hạn hơn như là vũ khí tự động chết người, các tác động đến thị trường lao động của tự động hóa, tội phạm mạng, quyền riêng tư hoặc xe tự lái. Đây không phải là những điều vô lý để một số người nghĩ đến, nhưng họ chủ yếu quan tâm đến các vấn đề tương đối khác so với những vấn đề mà AI cấp độ con người hoặc siêu trí tuệ có thể gây ra.

Bên cạnh đó, đúng là các vấn đề dài hạn trong cộng đồng nghiên cứu AI đã được chính thống hóa ở một mức độ nhất định. Nhưng sự chấp nhận đã đạt được vẫn chỉ là một phần rất nhỏ và, có thể hình dung, rất mong manh. Một số nhà nghiên cứu AI đã bắt đầu lo lắng rằng cuộc trao đổi công khai sẽ vượt khỏi tầm kiểm soát. Những hình ảnh vô nghĩa giống trong phim Terminator đang gây ra những ảnh hưởng xấu. Việc bôi nhọ một ngành nghiên cứu, một cộng đồng chuyên môn hay một sự nghiệp của ai đó không phải là một điều dễ chịu. Rất có khả năng, để phản ứng với một báo động giả về một đội quân robot xấu xa, cộng đồng AI sẽ tập hợp lại với tâm thế dè bủ đối với bất kỳ mối quan tâm thực sự nào về trí tuệ máy tiên tiến. Có thể

mọi thứ sẽ tốt nếu một phản ứng như vậy đặc biệt nhắm vào các thông tin truyền thông sai lệch. Tuy nhiên, những thiệt hại kèm theo cũng là điều đáng lo ngại. Nếu một chuẩn mực phát sinh trong đội ngũ nghiên cứu AI rằng việc nói về siêu trí tuệ hoặc tìm hiểu những rủi ro có thể xảy ra là thiếu khôn ngoan với nỗi lo sợ về việc sẽ “nổi giáo” cho các nhà phê bình, những kẻ sợ hãi, mất trí, hay những người có thể trở thành lãnh đạo, thì khi đó, những gì đạt được trong việc chính thống hóa nói trên sẽ sớm tan biến. Sau đó, chúng ta có thể bước vào một “mùa đông của vấn đề an toàn AI” mà trong đó những công việc được đề xuất trong cuốn sách này rất khó được thực hiện. Việc ngăn chặn sự phát triển của một động lực trái chiều như vậy là rất quan trọng. Con đường tốt nhất để phát triển siêu trí tuệ có lợi là con đường mà ở đó, các nhà phát triển AI cũng như các nhà nghiên cứu vấn đề an toàn AI là cùng chí hướng, và ở đó, ở một mức độ đáng kể, hai đối tượng này là một. Vì vậy, tôi kêu gọi tất cả các bên hãy kiên nhẫn, kiềm chế, cởi mở, và tham gia đối thoại và hợp tác trực tiếp bất cứ khi nào có thể.

Ở đây, tôi sẽ không tiện thể trả lời các bình luận đã đưa ra kể từ khi xuất bản sách. Tôi sẽ chỉ đưa ra một nhận xét nhỏ hướng tới những ai đã sở hữu phiên bản này và chủ yếu là những người bận rộn đến nỗi không còn đọc được những cuốn sách đã mua, ngoại trừ việc lướt qua mục lục và nội dung phần đầu và phần cuối: cụ thể, tôi chỉ ra rằng có nhiều yếu tố, ngoài tầm quan trọng vốn có của một vấn đề, có thể ảnh hưởng đến việc nó được bỏ công thảo luận trong bao nhiêu trang. Vì vậy, việc suy luận ra những gì mà tôi tin tưởng từ số trang sách là điều không thể. Chẳng hạn, tôi tập trung vào nguy cơ nhiều hơn là những triển vọng. Điều này không có nghĩa là tôi coi

nhẹ yếu tố thứ hai. Tôi chỉ tình cờ nghĩ rằng, tại thời điểm này trong lịch sử, trong khi chúng ta có thể lơ mơ nghĩ rằng có những điều tuyệt vời (và vô cùng lớn lao) để hy vọng nếu quá trình chuyển đổi trí tuệ diễn ra tốt đẹp, thì việc phát triển sự hiểu biết chi tiết chính xác về những điều cụ thể có thể tiến triển xấu để đảm bảo tránh được chúng là chuyện cấp thiết. Vì lẽ ấy, tôi dành nhiều trang để phân tích các kịch bản mà trong đó, một siêu trí tuệ duy nhất trở nên đủ mạnh để định hình tương lai theo các sở thích của nó; điều này không có nghĩa là tôi loại bỏ các kịch bản đa cực (chẳng hạn, xem Chương 11). Tôi thảo luận về mức độ khó khăn của vấn đề kiểm soát và làm thế nào các giải pháp tỏ ra hợp lý lại có thể thất bại; tuy nhiên, vấn đề kiểm soát cũng có thể trở nên dễ dàng không kém.

Tôi muốn một lần nữa cảm ơn tất cả những người đã giúp đỡ tôi hoàn thành cuốn sách này hoặc đã đóng góp cho việc đón nhận nó. Tôi cũng xin cảm ơn tất cả những ai cố gắng đóng góp với tinh thần xây dựng vào vấn đề mới mẻ này của toàn bộ chúng ta.

Nick Bostrom

Tháng 11 năm 2015

Chú thích

LỜI NÓI ĐẦU

1. Tuy nhiên không phải chú thích cuối sách nào cũng hữu ích.
2. Tôi không biết đó là điểm nào.

CHƯƠNG 1: SỰ PHÁT TRIỂN TRONG QUÁ KHỨ VÀ NHỮNG NĂNG LỰC HIỆN CÓ

1. Mức độ thu nhập đủ sống hiện nay là khoảng \$400 (Chen và Ravallion 2010). Một triệu khoản thu nhập đủ sống do đó bằng \$400.000.000. Tổng sản lượng kinh tế thế giới hiện tại là khoảng \$60.000.000.000.000 và trong những năm gần đây tăng trưởng với tốc độ hằng năm khoảng 4% (tỷ lệ tăng trưởng gộp hằng năm từ 1950, theo Maddison [2010]). Những con số này cho ra giá trị ước lượng trong sách, và tất nhiên đó chỉ là ước lượng gần đúng. Nếu trực tiếp nhìn vào số liệu về dân số, chúng ta sẽ thấy rằng hiện nay cần khoảng 1,5 tuần để dân số thế giới tăng thêm 1.000.000 người; nhưng như vậy là chưa đánh giá đủ tốc độ tăng trưởng của nền kinh tế vì thu thập theo đầu người cũng tăng. Năm 500 TCN, sau Cách mạng Nông nghiệp, dân số thế giới tăng một triệu người mỗi 200 năm – rất đáng kể vì tốc độ này của người tiền sử chỉ là một triệu người mỗi một triệu năm – như vậy đã có sự tăng tốc rất lớn xảy ra từ đó đến nay. Mặc dù vậy, điều gây ấn tượng mạnh là tổng lượng tăng trưởng kinh tế suốt 200 năm cách đây 7.000 năm giờ đây chỉ cần 90 phút, và

rằng tăng trưởng dân số thế giới trước đây cần hai thế kỷ giờ chỉ diễn ra trong một tuần rưỡi. Xem thêm Maddison (2005).

2. Sự tăng trưởng và tăng tốc kịch tính này có thể cho thấy việc xuất hiện một “điểm kỳ dị” đang đến, như đã được John von Neumann nhắc đến trong một cuộc trao đổi với nhà toán học Stanislaw Ulam:

Cuộc trao đổi của chúng tôi tập trung vào quá trình tăng tốc không ngừng của công nghệ và những thay đổi trong cơ chế sinh sống của con người, điều có thể liên quan đến sự xuất hiện của một điểm kỳ dị quan trọng nào đó đang đến gần trong lịch sử của cuộc đua mà khi vượt qua đó thì các hoạt động của con người mà ta từng biết, sẽ không thể tiếp tục. (Ulam 1958)

3. Hanson (2000).

4. Vinge (1993); Kurzweil (2005).

5. Sandberg (2010).

6. Van Zanden (2003); Maddison (1999, 2001); De Long (1998).

7. Hai phát biểu lạc quan thường được nhắc đi nhắc lại vào những năm 1960: “Trong vòng 20 năm, máy móc sẽ có khả năng thực hiện bất kỳ việc gì mà con người có thể” (Simon 1965, 96); “Trong vòng một thế hệ... vấn đề chế tạo trí tuệ nhân tạo về cơ bản sẽ được giải quyết” (Minsky 1967, 2). Xem bản tổng kết có hệ thống các dự đoán về AI trong Armstrong và Sotala (2012).

8. Ví dụ, xem Baum và cộng sự (2011) và Armstrong và Sotala (2012).

9. Tuy nhiên, điều đó có thể gợi ý rằng các nhà nghiên cứu AI biết ít về các dòng thời gian phát triển hơn họ nghĩ, nhưng điều này có

thể đúng trong cả hai trường hợp: họ có thể đánh giá quá cao hoặc đánh giá thấp thời gian đến AI.

10. Good (1965, 33).

11. Norbert Wiener, người đã e ngại về những hậu quả có thể xảy ra, là một ngoại lệ. Năm 1960, ông viết: “Nếu vì mục đích của mình mà chúng ta sử dụng một cỗ máy cơ học có sự vận hành mà chúng ta không thể can thiệp hiệu quả một khi đã khởi động nó, bởi hành động nhanh và không thể hủy bỏ đến mức chúng ta không có dữ liệu để can thiệp trước khi hành động hoàn tất, thì chúng ta nên chắc chắn rằng mục đích được giao cho cỗ máy chính là những gì chúng ta thực sự mong muốn chứ không phải chỉ là một sự bắt chước màu mè.” (Wiener 1960). Ed Fredkin đã bày tỏ những lo lắng của mình về AI siêu trí tuệ trong một cuộc phỏng vấn được mô tả trong McCorduck (1979). Đến năm 1970, bản thân Good đã viết về những rủi ro, và thậm chí kêu gọi thành lập một hiệp hội để đối phó với những mối nguy hiểm (Good [1970]; xem thêm bài báo [Good 1982] nơi ông báo trước một số ý tưởng về phương pháp “quy chuẩn gián tiếp” mà chúng ta sẽ thảo luận trong Chương 13). Đến năm 1984, Marvin Minsky cũng đã viết về nhiều mối lo chính yếu (Minsky 1984).

12. So sánh với Yudkowsky (2008a). Về tầm quan trọng của việc đánh giá hậu quả đạo đức của các công nghệ nguy hiểm tiềm tàng của tương lai trước khi chúng trở nên khả thi, xem Roache (2008).

13. McCorduck (1979).

14. Newell và cộng sự (1959).

15. Chương trình SAINTS, chương trình ANALOGY, và chương trình STUDENT (lần lượt). Xem Slagle (1963), Evans (1964, 1968), và Bobrow (1968).

16. Nilsson (1984).

17. Weizenbaum (1966).

18. Winograd (1972).

19. Cope (1996); Weizenbaum (1976); Moravec (1980); Thrun và cộng sự (2006); Buehler và cộng sự (2009); Koza và cộng sự (2003). Sở phương tiện cơ giới Nevada đã cấp giấy phép đầu tiên cho một chiếc xe không người lái vào tháng 5 năm 2012.

20. Hệ thống STANDUP (Ritchie và cộng sự 2007).

21. Schwartz (1987). Ở đây, Schwartz mô tả một quan điểm mang tính hoài nghi mà ông nghĩ rằng đã được thể hiện qua các bài viết của Hubert Dreyfus.

22. Một nhà phê bình thanh nhạc trong giai đoạn này là Hubert Dreyfus.

Những người hoài nghi nổi bật khác từ thời đại này bao gồm John Lucas, Roger Penrose và John Searle. Tuy nhiên, trong số này chỉ có Dreyfus chủ yếu quan tâm đến việc bác bỏ các tuyên bố về những thành tựu thực tế mà chúng ta nên mong đợi từ các mô hình hiện có trong AI (mặc dù anh ta dường như đã cởi mở với khả năng các mô hình mới có thể tiến xa hơn). Mục tiêu của Searle là các lý thuyết chức năng trong triết học tâm trí, chứ không phải sức mạnh công cụ của các hệ thống AI. Lucas và Penrose phủ nhận rằng một máy tính cổ điển có thể được lập trình để làm mọi thứ mà một nhà toán học người có thể làm, nhưng họ không phủ nhận rằng bất kỳ chức năng cụ thể nào về nguyên tắc có thể được tự động hóa hoặc AI cuối cùng có thể trở nên rất mạnh mẽ. Cicero nhận xét rằng “các triết gia chẳng bao giờ nói gì vô nghĩa” (Cicero 1923, 119); nhưng vẫn thật khó để nghĩ đến bất kỳ nhà tư tưởng quan trọng nào đã phủ

nhận tính khả thi của siêu trí tuệ của máy theo nghĩa được sử dụng trong cuốn sách này.

23. Tuy nhiên, đối với nhiều ứng dụng, việc học tập diễn ra trong mạng lưới thần kinh không khác mấy so với việc học diễn ra trong hồi quy tuyến tính, một kỹ thuật thống kê được phát triển bởi Adrien-Marie Legendre và Carl Friedrich Gauss vào đầu những năm 1800.

24. Thuật toán cơ bản được Arthur Bryson và Yu-Chi Ho mô tả là một phương pháp tối ưu hóa động nhiều giai đoạn vào năm 1969 (Bryson và Ho 1969). Ứng dụng vào mạng lưới thần kinh được đề xuất bởi Paul Werbos vào năm 1974 (Werbos 1994), nhưng chỉ sau khi David Rumelhart, Geoffrey Hinton và Ronald Williams làm việc vào năm 1986 (Rumelhart và cộng sự năm 1986) thì phương pháp này dần dần bắt đầu thấm vào nhận thức của một cộng đồng rộng lớn hơn.

25. Các mạng thiếu các lớp ẩn trước đây đã được chứng minh là có chức năng hạn chế nghiêm trọng (Minsky và Papert 1969).

26. VD: MacKay (2003).

27. Murphy (2012).

28. Chúng ta loại bỏ các chi tiết kỹ thuật khác nhau ở đây để không gây gánh nặng quá mức cho việc giải trình. Chúng ta sẽ có dịp xem lại một số vấn đề bị bỏ qua trong Chương 12 này.

29. Chương trình p là một mô tả của chuỗi x nếu p , chạy trên (một số cụ thể) máy Turing phổ dụng U , cho ra x ; chúng ta sẽ viết: $U(p) = x$. (Chuỗi x ở đây đại diện cho một thế giới khả thể.) Độ phức tạp Kolmogorov của x là $K(x) := \min \{ \ell(p) : U(p) = x \}$, trong đó $\ell(p)$ là độ dài của p tính bằng bit. Xác suất “Solomonoff” của x khi đó là:  0440_1, trong đó tổng được xác định trên toàn bộ các chương trình p (“tối thiểu”, nghĩa là không nhất thiết chia đôi), khi U cho ra một chuỗi bắt đầu bằng x (Hutter 2005).

30. Điều kiện Bayes trên chứng cứ E cho ra công thức:

 0440_2

(Xác suất của một mệnh đề [như E] là tổng xác suất của các thế giới khả thể mà trong đó nó là đúng.)

31. Hoặc chọn ngẫu nhiên một trong những hành động khả thể có độ thỏa dụng kỳ vọng cao nhất, trong trường hợp có nhiều hành động như vậy.

32. Chính xác hơn, độ thỏa dụng kỳ vọng của một hành động có thể được viết là  0440_3, trong đó tổng được tính trên tất cả các thế giới khả thể.

33. Ví dụ, xem Howson và Urbach (1993); Bernardo và Smith (1994); Russell và Norvig (2010).

34. Pearl (2009).

35. Wainwright và Jordan (2008). Các lĩnh vực ứng dụng của mạng Bayes là vô số; ví dụ, xem Pourret và cộng sự (2008).

36. Người ta có thể tự hỏi tại sao rất nhiều chi tiết được cung cấp cho trò chơi AI ở đây, mà một số có vẻ như là một khu vực ứng dụng

không quan trọng. Câu trả lời là chơi trò chơi có thể đưa ra một số con số rõ ràng nhất trong việc so sánh khả năng của con người và AI.

37. Newell và cộng sự (1958, 320).

38. Được cho là trong Vardi (2012).

39. Năm 1976, I. J. Good viết: “Một chương trình máy tính có sức mạnh Grandmaster sẽ mang lại cho chúng ta trong một loạt [trí tuệ máy siêu thông minh]” (Good 1976). Năm 1979, Douglas Hofstadter viết trong cuốn sách đoạt giải Pulitzer của ông tên là Gödel, Escher, Bach: “Câu hỏi: Sẽ có những chương trình cờ vua có thể đánh bại bất cứ ai? Suy đoán: Không. Có thể có những chương trình có thể đánh bại bất cứ ai trong cờ vua, nhưng chúng sẽ không phải là chương trình chỉ chơi cờ vua. Chúng sẽ là những chương trình trí tuệ tổng quát, và chúng sẽ có khí chất như con người. ‘Bạn có muốn chơi cờ vua không?’ ‘Không, tôi chán cờ vua rồi. Hãy nói về thơ đi.’” (Hofstadter [1979] 1999, 678).

40. Thuật toán là tìm kiếm minimax với cắt tỉa alpha-beta, được sử dụng với chức năng đánh giá tự nghiệm trên các thế cờ của môn cờ vua. Kết hợp với một thư viện thế cờ khai cuộc và tàn cuộc phong phú cùng nhiều chiến thuật khác, điều này có thể tạo ra một trình chơi cờ có năng lực.

41. Mặc dù những tiến bộ gần đây chỉ tập trung vào việc học quá trình đánh giá tự nghiệm từ các trò chơi được mô phỏng, nhiều thuật toán cơ bản có thể cũng sẽ hoạt động tốt cho nhiều trò chơi khác.

42. Samuel (1959); Schaeffer (1997, chương 6).

43. Schaeffer và cộng sự (2007).

44. Berliner (1980a, b).

45. Tesauro (1995).
46. Bao gồm GNU (xem Silver [2006]) và Snowie (xem Gammoned.net [2012]).
47. Chính Lenat đã giúp một tay trong việc hướng dẫn quy trình thiết kế hạm đội. Ông đã viết: “Do đó, tỷ lệ chiến thắng cuối cùng phải là khoảng 60/40% Lenat/Eurisko, mặc dù điểm quan trọng ở đây là không bên nào có thể giành chiến thắng một mình” (Lenat 1983, 80).
48. Lenat (1982, 1983).
49. Cirasella và Kopec (2006).
50. Kasparov (1996, 55).
51. Newborn (2011).
52. Keim và cộng sự (1999).
53. Xem Armstrong (2012).
54. Sheppard (2002).
55. Wikipedia (2012a).
56. Markoff (2011).
57. Rubin và Watson (2011).
58. Elyasaf và cộng sự (2011).
59. KGS (2012).
60. Nilsson (2009, 318). Knuth chắc chắn đã nói quá quan điểm của mình.
- Có rất nhiều “tác vụ tư duy” mà AI đã không thực hiện thành công, như việc phát minh ra một lĩnh vực mới của toán học thuần túy, thực hiện một kiểu triết lý, viết một cuốn tiểu thuyết trinh thám tuyệt vời, cầm đầu một cuộc đảo chính hay thiết kế một sản phẩm tiêu dùng chủ chốt mới.
61. Shapiro (1992).

62. Người ta có thể suy đoán rằng một lý do cho việc rất khó để đạt được khả năng như con người về nhận thức, điều khiển vận động, hiểu lễ thường và hiểu ngôn ngữ là bộ não của chúng ta dành riêng cho các chức năng này (cấu trúc thần kinh được tối ưu hóa theo thời gian tiến hóa). Ngược lại, suy nghĩ logic và kỹ năng như chơi cờ vua không phải là tự nhiên đối với chúng ta; vì vậy có lẽ chúng ta buộc phải dựa vào một nguồn tài nguyên với mục đích phổ thông để thực hiện các nhiệm vụ này. Có lẽ những gì bộ não làm khi chúng ta tham gia vào lý luận hoặc tính toán logic rõ ràng theo một cách nào đó tương tự việc vận hành một “cỗ máy ảo”, một bản mô phỏng tinh thần chậm và cồng kềnh của một máy tính đa năng. Sau đó, người ta có thể nói (một cách hơi huyền ảo) rằng một chương trình AI cổ điển không giả lập quá nhiều suy nghĩ của con người như một người suy nghĩ logic đang “giả lập” một chương trình AI.

63. Ví dụ này gây tranh cãi: một quan điểm thiểu số, đại diện cho khoảng 20% người trưởng thành ở Mỹ và số lượng tương tự ở nhiều quốc gia phát triển khác, cho rằng Mặt trời xoay quanh Trái đất (Crabtree 1999; Dean 2005).

64. World Robotics (2011).

65. Ước tính từ dữ liệu trong Guizzo (2010).

66. Holley (2009).

67. Phương pháp thống kê hỗn hợp dựa trên quy tắc cũng được sử dụng, nhưng chúng hiện chỉ là một phần nhỏ của bức tranh.

68. Cross và Walker (1994); Hedberg (2002).

69. Dựa trên số liệu thống kê từ TABB Group, một công ty nghiên cứu thị trường vốn có trụ sở tại New York và London (trao đổi cá nhân).

70. CFTC và SEC (2010). Để có góc nhìn khác về các sự kiện ngày 6 tháng 5 năm 2010, hãy xem CME Group (2010).

71. Trong văn bản không có yếu tố nào được hiểu là một đối số chống lại giao dịch thuật toán tần suất cao, thường có thể thực hiện một chức năng có lợi bằng cách tăng thanh khoản và hiệu quả thị trường.

72. Một vụ mua tháo thị trường nhỏ hơn đã xảy ra vào ngày 1 tháng 8 năm 2012, một phần là do công cụ ngắt mạch của Google cũng không được lập trình để tạm dừng giao dịch nếu có sự thay đổi lớn về số lượng cổ phiếu được giao dịch (Popper 2012). Điều này một lần nữa báo trước một chủ đề khác sau này: khó khăn trong việc dự đoán tất cả các cách cụ thể trong đó một số quy tắc có vẻ đặc biệt hợp lý lại có thể sai.

73. Nilsson (2009, 319).

74. Minsky (2006); McCarthy (2007); Beal và Winston (2009).

75. Peter Norvig, trao đổi cá nhân. Các lớp học máy cũng rất phổ biến, phản ánh một làn sóng cường điệu có phần trực giao của “big data” (lấy cảm hứng từ Google và Giải thưởng Netflix).

76. Armstrong và Sotala (2012).

77. Müller và Bostrom (2016).

78. Xem Baum và cộng sự (2011), một khảo sát khác được trích dẫn trong đó, và trong Sandberg và Bostrom (2011).

79. Nilsson (2009).

80. Một lần nữa, điều kiện là không xảy ra thảm họa phá huỷ nền văn minh.

Định nghĩa của HLMI được Nilsson sử dụng là “AI có thể thực hiện khoảng 80% công việc tương đương hoặc tốt hơn con người”

(Kruel 2012).

81. Bảng này cho thấy kết quả của bốn cuộc thăm dò khác nhau cũng như kết quả tổng hợp. Hai cuộc thăm dò đầu tiên được thực hiện tại các hội nghị học thuật: PT-AI, những người tham gia hội nghị Triết học và Lý thuyết về AI ở Thessaloniki năm 2011 (được hỏi vào tháng 11 năm 2012), với tỷ lệ trả lời là 43 trên 88; và AGI, những người tham gia hội nghị Trí tuệ Nhân tạo Tổng thể và Tác động và Rủi ro của Trí tuệ Nhân tạo Tổng thể ở Oxford, tháng 12 năm 2012 (tỷ lệ trả lời: 72/111). Cuộc thăm dò của EETN đã lấy mẫu các thành viên của Hiệp hội Trí tuệ nhân tạo Hy Lạp, một tổ chức chuyên nghiệp gồm các nhà nghiên cứu đã công bố trong lĩnh vực, vào tháng 4 năm 2013 (tỷ lệ trả lời: 26/250). Cuộc thăm dò TOP100 đã lấy ý kiến 100 tác giả hàng đầu về trí tuệ nhân tạo được đo bằng tỷ lệ trích dẫn, vào tháng 5 năm 2013 (tỷ lệ trả lời: 29/100).

82. Các cuộc phỏng vấn với khoảng 28 người (tại thời điểm viết bài) trong lĩnh vực AI và các chuyên gia liên quan đã được Kruel (2011) công bố.

83. Biểu đồ cho thấy các giá trị ước tính trung bình tái chuẩn hóa. Giá trị trung bình là khác nhau đáng kể. Ví dụ: các ước tính trung bình cho kết quả “rất xấu” của Drake là 7,6% (cho TOP100) và 17,2% (cho nhóm đánh giá chuyên gia kết hợp).

84. Có một tài liệu đáng chú ý chứng minh sự không đáng tin cậy của các dự báo chuyên gia trong nhiều lĩnh vực, và có mọi lý do để nghĩ rằng nhiều phát hiện trong nghiên cứu này cũng áp dụng cho lĩnh vực trí tuệ nhân tạo. Cụ thể, các nhà dự báo có xu hướng quá tự tin vào dự đoán của bản thân, tin rằng mình chính xác hơn so với thực tế nên thường gán quá ít xác suất cho khả năng giả thuyết

được ưa thích nhất của họ là sai (Tetlock 2005). (Nhiều thiên kiến khác cũng đã được ghi lại; ví dụ, xem Gilovich và cộng sự [2002].) Tuy nhiên, sự không chắc chắn là một thực tế tất yếu về tình trạng của con người và nhiều hành động của chúng ta không thể tránh khỏi việc dựa vào những kỳ vọng về việc hậu quả lâu dài nào là khả thể hơn: nói cách khác, về dự đoán xác suất. Từ chối đưa ra dự đoán xác suất rõ ràng sẽ không làm cho vấn đề nhận thức biến mất mà chỉ che giấu nó (Bostrom 2007). Thay vào đó, chúng ta nên phản hồi bằng chứng về sự tự tin thái quá bằng cách mở rộng khoảng tự tin (hoặc “khoảng tin cậy”) của mình (nghĩa là, bằng cách mở rộng các hàm tín nhiệm của chúng ta) và nói chung, chúng ta phải đấu tranh hết sức có thể với những thành kiến của mình, bằng cách xem xét các quan điểm khác nhau và hướng tới sự trung thực trí tuệ. Về lâu dài, chúng ta cũng có thể nghiên cứu phát triển các kỹ thuật, phương pháp đào tạo và các tổ chức có thể giúp hiệu chuẩn tốt hơn. Xem thêm Armstrong và Sotala (2012).

CHƯƠNG 2: CON ĐƯỜNG ĐẾN VỚI SIÊU TRÍ TUỆ

1. Giống với định nghĩa trong Bostrom (2003c) và Bostrom (2006a). Nó cũng có thể được so sánh với định nghĩa của Shane Legg (Trí tuệ đo lường khả năng của tác tử để đạt được mục tiêu trong một loạt các môi trường khác nhau) và các dạng hình thức hóa của định nghĩa này (Legg 2008). Nó cũng rất giống với định nghĩa của Good về cỗ máy siêu thông minh trong Chương 1 (“một cỗ máy có thể vượt xa mọi hoạt động trí tuệ của bất kỳ người nào dù thông minh đến đâu”).

2. Vì lý do tương tự, chúng tôi không đưa ra giả định nào về việc liệu một cỗ máy siêu trí tuệ có thể có ý định thực sự hay không (xin

lỗi Searle, nó có thể; nhưng điều này có vẻ không liên quan đến những lo ngại của cuốn sách này). Và chúng ta không có vị trí nào trong cuộc tranh luận thuyết nội tại/thuyết ngoại tại về nội dung tinh thần đang hoành hành trong các tài liệu triết học, hoặc về vấn đề liên quan của luận đề tư duy mở rộng (Clark và Chalmers 1998).

3. Turing (1950, 456).

4. Turing (1950, 456).

5. Chalmers (2010); Moravec (1976, 1988, 1998, 1999).

6. Xem Moravec (1976). Một lập luận tương tự được đưa ra bởi David Chalmers (2010).

7. Xem thêm Shulman và Bostrom (2012), trong đó các vấn đề này được bàn chi tiết hơn.

8. Legg (2008) đưa ra lý do này để hỗ trợ cho tuyên bố rằng con người sẽ có thể tóm tắt lại tiến trình tiến hóa trong khoảng thời gian ngắn hơn nhiều và với tài nguyên tính toán giảm (trong khi lưu ý rằng tài nguyên tính toán không được điều chỉnh của tiến hóa là xa tầm với). Baum (2004) lập luận rằng một số phát triển liên quan đến AI đã xảy ra trước đó, với bản thân tổ chức của bộ gen thể hiện một phép biểu diễn có giá trị cho các thuật toán tiến hóa.

9. Whitman và cộng sự (1998); Sabrosky (1952).

10. Schultz (2000).

11. Menzel và Giurfa (2001, 62); Truman và cộng sự (1993).

12. Sandberg và Bostrom (2008).

13. Xem Legg (2008), thảo luận thêm về điểm này và về sự hứa hẹn của các hàm hoặc môi trường xác định sự phù hợp dựa trên bối cảnh phẳng lạng của các bài kiểm tra trí tuệ thuần túy.

14. Xem Bostrom và Sandberg (2009b), một phân loại và thảo luận chi tiết hơn về các cách thức mà các kỹ sư có thể vượt trội hơn so với chọn lọc tiến hóa lịch sử.

15. Phân tích đã nhắm tới các hệ thống thần kinh của các sinh vật sống, mà không cần tham khảo chi phí mô phỏng các thực thể hoặc môi trường ảo xung quanh như là một phần của hàm thích nghi. Điều hợp lý là một hàm thích nghi đầy đủ có thể kiểm tra năng lực của một sinh vật cụ thể trong các hoạt động ít hơn nhiều so với việc mô phỏng tất cả các tính toán thần kinh của não bộ sinh vật đó trong suốt vòng đời tự nhiên của nó. Các chương trình AI ngày nay thường phát triển và hoạt động trong môi trường rất trừu tượng (các trình chứng minh định lý trong thế giới toán học biểu tượng, tác tử trong thế giới trò chơi,...).

Một người hoài nghi có thể khẳng định rằng một môi trường trừu tượng sẽ không đủ cho sự tiến hóa của trí tuệ nói chung, tin rằng môi trường ảo sẽ cần gần giống với môi trường sinh học thực tế mà tổ tiên chúng ta phát triển. Thực tế, tạo ra một thế giới ảo vật lý sẽ đòi hỏi một sự đầu tư lớn hơn nhiều vào tài nguyên tính toán so với bản mô phỏng của một thế giới đồ chơi đơn giản hoặc một miền vấn đề trừu tượng (trong khi tiến hóa có quyền truy cập tự do vào một thế giới thực vật lý). Trong trường hợp giới hạn, nếu độ chính xác vi mô hoàn toàn được nhấn mạnh, các yêu cầu tính toán sẽ tăng lên đến một tỷ lệ lố bịch. Tuy nhiên, sự bi quan cực đoan như vậy gần như chắc chắn là không có cơ sở; dường như môi trường tốt nhất để phát triển trí tuệ là một môi trường bất chước tự nhiên gần giống như có thể nhìn thấy được. Trái lại, rất có thể sử dụng môi trường chọn lọc nhân tạo sẽ hiệu quả hơn và không giống với môi trường của tổ

tiên của chúng ta, một môi trường được thiết kế đặc biệt để thúc đẩy sự thích nghi làm tăng cường loại trí tuệ mà chúng ta đang tìm cách phát triển (như lý luận trừu tượng và các kỹ năng giải quyết vấn đề chung, trái ngược với các phản ứng bản năng nhanh nhất hoặc một hệ thống hình ảnh được tối ưu hóa cao).

16. Wikipedia (2012b).

17. Xem thông tin chung về lý thuyết chọn lọc quan sát trong Bostrom (2002a). Xem ứng dụng cụ thể cho vấn đề hiện tại trong Shulman và Bostrom (2012). Xem dẫn nhập ngắn trong Bostrom (2008b).

18. Sutton và Barto (1998, 21f); Schultz và cộng sự (1997).

19. Đây là thuật ngữ của Eliezer Yudkowsky; ví dụ, xem Yudkowsky (2007).

20. Đây là kịch bản được mô tả bởi Good (1965) và Yudkowsky (2007). Tuy nhiên, người ta cũng có thể xem xét một giải pháp thay thế trong đó trình tự lặp có một số bước không liên quan đến tăng cường trí tuệ mà thay vào đó là đơn giản hóa thiết kế. Nghĩa là, ở một số giai đoạn, AI hạt giống có thể tự viết lại để làm cho những cải tiến tiếp theo dễ tìm thấy hơn.

21. Helmstaedter và cộng sự (2011).

22. Andres và cộng sự (2012).

23. Điều này thích hợp để cho phép các hình thức hữu ích của chức năng nhận thức và giao tiếp; nhưng vẫn rất nghèo nàn so với giao diện được cung cấp bởi các cơ và cơ quan cảm giác của một cơ thể người bình thường.

24. Sandberg (2013).

25. Xem phần “Computer requirements” (Các yêu cầu về máy tính) trong Sandberg và Bostrom (2008, 79–81).

26. Mức độ thành công thấp hơn có thể là một mô phỏng não có động lực học vi mô gợi ý về mặt sinh học và hiển thị hàng loạt các hoạt động đặc thù loài mới nổi như trạng thái ngủ chậm hoặc tính linh hoạt phụ thuộc vào hoạt động. Trong khi một bản mô phỏng như vậy có thể là một thử nghiệm hữu ích cho nghiên cứu khoa học thần kinh (mặc dù có thể làm phát sinh các vấn đề đạo đức nghiêm trọng), nó sẽ không được tính là giả lập hoàn chỉnh não bộ, trừ khi bản mô phỏng đủ chính xác để có thể thực hiện một phần đáng kể công việc trí tuệ mà bộ não được mô phỏng có khả năng. Theo nguyên tắc thông thường, chúng ta có thể nói rằng để bản mô phỏng não người được tính là giả lập hoàn chỉnh não bộ, cần phải có khả năng diễn đạt những suy nghĩ bằng lời nói mạch lạc hoặc có khả năng học cách làm như vậy.

27. Sandberg and Bostrom (2008).

28. Sandberg and Bostrom (2008). Xem giải thích thêm có thể được tìm thấy trong báo cáo gốc.

29. Sơ đồ đầu tiên được mô tả trong Albertson và Thomson (1976) và White và cộng sự (1986). Mạng lưới kết hợp (và trong một số trường hợp đã được chỉnh sửa) có thể xem trên trang web của “WormAtlas” (<http://www.wormatlas.org/>).

30. Xem về những nỗ lực giả lập *C. elegans* và kết quả của những nỗ lực này trong Kaufman (2011). Kaufman trích dẫn một nghiên cứu sinh tiến sĩ đầy tham vọng làm việc trong ngành tên là David Dalrymple rằng “Với các kỹ thuật quang di truyền, chúng tôi vừa đạt tới một điểm mà ở đó việc đạt được khả năng đọc và ghi vào

bất cứ nơi nào trong hệ thống thần kinh của một con *C. elegans* sống, sử dụng hệ thống tự động công suất cao không phải là một đề xuất thái quá... Tôi mong đợi hoàn thành *C. Elegans* trong từ hai đến ba năm. Tôi sẽ vô cùng ngạc nhiên nếu đây vẫn là một vấn đề bỏ ngỏ vào năm 2020, dù có thể nào đi nữa” (Dalrymple 2011). Các mô hình não nhằm đến chủ nghĩa hiện thực sinh học được mã hóa bằng tay (chứ không phải được tạo tự động) đã đạt được một số chức năng cơ bản; ví dụ, xem Eliasmith và cộng sự (2012).

31. *Caenorhabditis elegans* có một số tính chất đặc biệt thuận tiện. Ví dụ, sinh vật này trong suốt, và kết nối trong hệ thần kinh của nó không thay đổi giữa các cá thể.

32. Nếu AI hình thái thần kinh chứ không phải giả lập hoàn chỉnh não bộ mới là sản phẩm cuối cùng, thì có thể hoặc không thể xảy ra trường hợp những hiểu biết có liên quan sẽ phát sinh thông qua các nỗ lực mô phỏng bộ não của con người. Có thể hình dung rằng một số chức năng vỏ não quan trọng sẽ được phát hiện trong quá trình nghiên cứu về bộ não động vật (không phải con người). Một số bộ não động vật có thể dễ nghiên cứu hơn bộ não người và bộ não nhỏ hơn sẽ cần ít tài nguyên hơn để quét và mô hình hóa. Nghiên cứu về bộ não động vật cũng sẽ chịu ít quy định hơn. Thậm chí có thể hình dung rằng trí tuệ máy cấp độ con người đầu tiên sẽ được tạo ra bằng cách hoàn thành một bản giả lập hoàn chỉnh não bộ của một số động vật phù hợp và sau đó tìm cách tăng cường trí tuệ kỹ thuật số được tạo ra. Do đó, loài người có thể có được nó từ một con chuột hoặc khỉ thí nghiệm được nâng cấp.

33. Uauy và Dangour (2006); Georgieff (2007); Stewart và cộng sự (2008); Eppig và cộng sự (2010); Cotman và Berchtold (2002).

34. Theo Tổ chức Y tế Thế giới năm 2007, gần hai tỷ người không đủ lượng iốt (The Lancet 2008). Thiếu iốt nghiêm trọng cản trở sự phát triển thần kinh và dẫn đến đần độn, liên quan đến việc mất trung bình khoảng 12,5 điểm IQ (Qian và cộng sự 2005). Tình trạng này có thể được ngăn chặn dễ dàng và không tốn kém thông qua việc bổ sung iốt vào muối (Horton và cộng sự 2008).

35. Bostrom và Sandberg (2009a).

36. Bostrom và Sandberg (2009b). Sự gia tăng hiệu suất giả định điển hình từ tăng cường dược lý và dinh dưỡng nằm trong khoảng 10-20% đối với các tác vụ thử nghiệm đo lường bộ nhớ làm việc, sự chú ý, ... và chỉ ra kết quả được cải thiện tương ứng trong các tình huống giải quyết vấn đề trong thế giới thực (Repantis và cộng sự 2010). Ví dụ, trong một số trường hợp, có thể có sự suy giảm bù cho một số kích thước hiệu suất không được đo bằng các tác vụ thử nghiệm (Sandberg và Bostrom 2006).

37. Nếu có một cách dễ dàng để tăng cường nhận thức, người ta sẽ mong đợi tiến hóa đã tận dụng được nó. Do đó, loại thuốc thông minh hứa hẹn nhất cần nghiên cứu có thể là một loại hứa hẹn sẽ tăng cường trí tuệ theo một cách nào đó mà chúng ta có thể thấy là sẽ ít phù hợp hơn trong môi trường của tổ tiên chúng ta, ví dụ, bằng cách tăng kích thước đầu khi mới sinh hoặc tăng chuyển hóa glucose của não. Xem thảo luận chi tiết hơn về ý tưởng này (cùng với một số bằng cấp quan trọng) trong Bostrom (2009b).

38. Tinh trùng khó sàng lọc hơn vì trái ngược với phôi, chúng chỉ bao gồm một tế bào và một tế bào cần phải bị phá hủy để thực hiện giải trình tự. Tế bào trứng cũng chỉ bao gồm một tế bào; tuy nhiên, sự phân chia tế bào thứ nhất và thứ hai không đối xứng và tạo ra một

tế bào con với rất ít tế bào chất, hay một thể cực. Vì các thể cực có bộ gen giống như tế bào chính và là dư thừa (cuối cùng sẽ thoái hóa), chúng có thể được sinh thiết và sử dụng để sàng lọc (Gianaroli 2000).

39. Mỗi phương pháp này đều phải chịu một số tranh cãi về đạo đức khi ra mắt, nhưng dường như sự chấp nhận thường sẽ tăng lên. Thái độ đối với kỹ thuật di truyền trên con người và chọn lọc phôi khác nhau đáng kể giữa các nền văn hóa, cho thấy rằng sự phát triển và ứng dụng các kỹ thuật mới có thể sẽ diễn ra ngay cả khi một số quốc gia ban đầu chấp nhận lập trường thận trọng, mặc dù tốc độ mà việc này xảy ra sẽ bị ảnh hưởng bởi đạo đức, tôn giáo và áp lực chính trị.

40. Davies và cộng sự (2011); Benyamin và cộng sự (2013); Plomin và cộng sự (2013). Xem thêm Mardis (2011); Hsu (2012).

41. Khả năng di truyền theo nghĩa rộng của IQ người trưởng thành thường được ước tính trong khoảng 0,5-0,8 trong tầng lớp trung lưu của các quốc gia phát triển (Bouchard 2004, 148). Khả năng di truyền theo nghĩa hẹp, đo tỷ lệ phương sai của các yếu tố di truyền cộng hợp, thì thấp hơn (trong khoảng 0,3-0,5) nhưng vẫn đáng kể (Devlin và cộng sự 1997; Davies và cộng sự 2011; Visscher và cộng sự 2008). Những ước tính này có thể khác nhau giữa các quần thể và môi trường khác nhau, vì khả năng di truyền khác nhau tùy thuộc vào quần thể và môi trường được nghiên cứu. Ví dụ, khả năng di truyền thấp hơn đã được tìm thấy ở trẻ em đến từ môi trường thiếu thốn (Benyamin và cộng sự 2013; Turkheimer và cộng sự 2003). Nisbett và cộng sự (2012) xem xét nhiều ảnh hưởng khác nhau của môi trường đến sự thay đổi trong khả năng nhận thức.

42. Một số đoạn tiếp sau đây tập trung nhiều vào công trình cùng nghiên cứu với Carl Shulman (Shulman và Bostrom 2014).

43. Bảng này được trích từ Shulman và Bostrom (2014). Nó dựa trên một mô hình mô phỏng giả định một phân phối Gauss của IQ dự đoán giữa các phôi thai với độ lệch chuẩn là 7,5 điểm. Lượng tăng cường nhận thức thu được với số lượng phôi thai khác nhau phụ thuộc vào mức độ khác nhau giữa các phôi thai trong các biến thể di truyền cộng hợp với hiệu ứng đã biết. Anh chị em có hệ số liên quan là $1/2$ và các biến thể di truyền cộng hợp phổ biến bằng hoặc thấp hơn 0,5 so với phương sai của trí tuệ mềm ở người trưởng thành (Davies và cộng sự 2011). Hai thông tin trên cho thấy độ lệch chuẩn dân số quan sát được ở các nước phát triển là 15 điểm, độ lệch chuẩn của ảnh hưởng di truyền trong một lô phôi sẽ là 7,5 điểm hoặc ít hơn.

44. Với thông tin không hoàn hảo về các hiệu ứng di truyền cộng hợp về khả năng nhận thức, quy mô hiệu ứng sẽ bị giảm. Tuy nhiên, ngay cả một lượng nhỏ kiến thức cũng sẽ thành công tương đối, vì lợi ích từ việc chọn lọc không tăng tuyến tính với tỷ lệ phương sai mà chúng ta có thể dự đoán. Thay vào đó, hiệu quả của quá trình chọn lọc phụ thuộc vào độ lệch chuẩn của IQ trung bình được dự đoán, tăng theo căn bậc hai của phương sai. Ví dụ: nếu một người có thể chiếm 12,5% phương sai, thì điều này có thể mang lại hiệu ứng bằng một nửa so với những gì nêu trong Bảng 1 (đang giả định là 50%). Để so sánh, một nghiên cứu gần đây (Rietveld và cộng sự 2013) tuyên bố đã xác định được 2,5% phương sai.

45. Để so sánh, các phương pháp tiêu chuẩn ngày nay tạo ra ít hơn 10 phôi.

46. Tế bào gốc từ người trưởng thành và phôi có thể được nuôi để phát triển thành tế bào tinh trùng và trứng, sau đó có thể được hợp nhất để tạo ra phôi (Nagy và cộng sự 2008; Nagy và Chang 2007). Tế bào tiền thân của trứng cũng có thể hình thành phôi nang đơn tính, (phôi không thụ tinh và không thể phát triển) có thể tạo ra các dòng tế bào gốc phôi cho quá trình (Mai và cộng sự 2007).

47. Ý kiến này là của Katsuhiko Hayashi, được nêu trong Cyranoski (2013).

Nhóm Hinxton, một nhóm quốc tế gồm các nhà khoa học thảo luận về vấn đề đạo đức và những khó khăn liên quan đến tế bào gốc, dự đoán vào năm 2008 rằng các loại giao tử có nguồn gốc từ tế bào người sẽ có sẵn trong 10 năm tới (Nhóm Hinxton 2008), và vẫn phù hợp với diễn tiến hiện thời.

48. Sparrow (2013); Miller (2012); *The Uncertain Future* (tạm dịch: *Tương lai bất định*) (2012).

49. Sparrow (2013).

50. Những nỗi lo muôn thuở có thể tập trung vào các tác động đã dự đoán đối với sự bất bình đẳng xã hội, tính an toàn y tế của phương pháp, lo ngại về một cuộc chạy đua tăng cường vô nghĩa, quyền và trách nhiệm của cha mẹ đối với con cái tương lai của họ, bóng tối ưu sinh thế kỷ 20, khái niệm về phẩm giá con người và giới hạn tham gia phù hợp của các quốc gia trong việc lựa chọn sinh sản của công dân. (Xem thảo luận về vấn đề đạo đức liên quan đến tăng cường nhận thức trong Bostrom và Ord [2006], Bostrom và Roache [2011], và Sandberg và Savulescu [2011].) Một số dòng tôn giáo có thể mang đến thêm các mối quan ngại tập trung vào vị thế đạo đức

của phôi thai hoặc những giới hạn của con người trong khuôn khổ tạo hóa.

51. Để ngăn chặn các tác động tiêu cực của giao phối cận huyết, việc chọn lọc phôi lặp lại sẽ đòi hỏi một nguồn hiến tặng ban đầu lớn hoặc năng lực chọn lọc đáng kể để giảm các allele lặn có hại. Hoặc là sự thay thế sẽ có xu hướng thúc đẩy con cái ít liên quan chặt chẽ về mặt di truyền với bố mẹ chúng (và liên quan nhiều hơn với nhau).

52. Chúng ta vẫn chưa biết liệu di truyền học biểu sinh sẽ là trở ngại khó khăn ra sao (Chason và cộng sự 2011; Iliadou và cộng sự 2011).

53. Mặc dù khả năng nhận thức là một đặc điểm khá dễ di truyền, có thể có ít hoặc không có các allele hoặc hiện tượng đa hình phổ biến nào có ảnh hưởng tích cực lớn đến trí tuệ (Davis và cộng sự 2010; Davies và cộng sự

2011; Rietveld và cộng sự 2013). Khi các phương pháp giải trình tự được cải thiện, việc khớp các allele tần suất thấp với các đặc điểm tương quan về nhận thức và hành vi của chúng sẽ ngày càng khả thi. Có một số bằng chứng lý thuyết cho thấy rằng một số allele gây ra rối loạn di truyền ở người đồng hợp tử có thể mang lại lợi thế nhận thức rõ rệt ở người mang gen dị hợp tử, dẫn đến dự đoán rằng các nhóm dị hợp tử của các bệnh Gaucher, Tay-Sachs và Niemann-Pick sẽ cao hơn 5 điểm so với nhóm đối chứng (Cochran và cộng sự 2006). Thời gian sẽ cho biết điều này là đúng hay sai.

54. Một bài viết (Nachman và Crowell 2000) ước tính 175 đột biến trên mỗi bộ gen mỗi thế hệ. Một công trình khác (Lynch 2010), bằng các phương pháp khác nhau, ước tính rằng trẻ sơ sinh trung bình có từ 50 đến 100 đột biến mới, còn Kong và cộng sự (2012) đưa ra hàm

ý về một con số khoảng 77 đột biến mới mỗi thế hệ. Hầu hết các đột biến này không ảnh hưởng đến chức năng, hoặc chỉ ảnh hưởng ở mức độ nhẹ không thể nhận ra; nhưng các tác động kết hợp của nhiều đột biến rất nhẹ có thể gây suy giảm đáng kể về thể lực. Xem thêm Crow (2000).

55. Crow (2000); Lynch (2010).

56. Tương tự trong Shulman và Bostrom (2014).

57. Bostrom (2008b).

58. Có một số cảnh báo quan trọng cho ý tưởng này. Có thể bộ gen thường thấy (điển hình) sẽ cần một số điều chỉnh để tránh các vấn đề. Ví dụ, một số phần của bộ gen có thể được điều chỉnh để tương tác với các phần khác theo giả định rằng tất cả các phần hoạt động với một mức độ hiệu quả nhất định. Việc tăng hiệu quả của những phần đó có thể dẫn đến tình trạng quá tải trên một số con đường trao đổi chất.

59. Những khuôn mặt này được Mike tạo ra từ những bức ảnh chân dung được chụp bởi Virtual Flavius (Mike 2013).

60. Tất nhiên, chúng có thể có một số hiệu ứng sớm hơn, bằng cách thay đổi kỳ vọng của mọi người về những gì sắp xảy ra.

61. Louis Harris và cộng sự (1969); Mason (2003).

62. Kalfoglou và cộng sự (2004).

63. Dữ liệu rõ ràng là hạn chế, nhưng trong các nghiên cứu lâu dài (longitudinal), các cá nhân được chọn cho kết quả 1 trên 10.000 trong các bài kiểm tra khả năng thời thơ ấu đã được chứng minh rằng về cơ bản có nhiều khả năng trở thành giáo sư, lấy bằng sáng chế và thành công trong kinh doanh hơn so với những người có chút điểm số kém xuất chúng hơn một chút (Kell và cộng sự 2013). Roe

(1953) đã nghiên cứu 64 nhà khoa học xuất chúng và tìm thấy khả năng nhận thức trung bình cao hơn từ ba đến bốn độ lệch chuẩn so với quy chuẩn của nhóm và cao hơn đáng kể so với các nhà khoa học nói chung. (Khả năng nhận thức cũng tương quan với thu nhập trọn đời và với các kết quả phi tài chính như tuổi thọ, tỷ lệ ly hôn và khả năng bỏ học [Deary 2012].) Sự thay đổi tăng dần của phân phối khả năng nhận thức sẽ có tác động lớn không tương xứng ở phần cuối, đặc biệt là tăng số lượng người có năng khiếu cao và giảm số người chậm phát triển và không có khả năng học tập. Xem thêm Bostrom và Ord (2006) và Sandberg và Savulescu (2011).

64. Ví dụ: Warwick (2002). Stephen Hawking thậm chí còn cho rằng thực hiện bước này có thể là cần thiết để theo kịp những tiến bộ về trí tuệ máy: “Chúng ta phải phát triển càng nhanh càng tốt các công nghệ có thể kết nối trực tiếp giữa não và máy tính, để bộ não nhân tạo đóng góp cho trí tuệ của con người thay vì chống lại nó” (ghi lại trong Walsh [2001]). Ray Kurzweil cũng đồng tình với ý kiến này: “Theo như khuyến nghị của Hawking liên quan đến kết nối trực tiếp giữa não và máy tính, tôi đồng ý rằng điều này vừa hợp lý, vừa đáng mong đợi và vừa không thể tránh khỏi. [nguyên văn] Đây đã là đề xuất của tôi trong nhiều năm” (Kurzweil 2001).

65. Xem Lebedev và Nicolelis (2006); Birbaumer và cộng sự (2008); Mak và Wolpaw (2009); và Nicolelis và Lebedev (2009). Một quan điểm mang tính cá nhân hơn về vấn đề tăng cường thông qua cấy ghép có thể được tìm thấy trong Chorost (2005, Chap 11).

66. Smeding và cộng sự (2006).

67. Degnan và cộng sự (2002).

68. Dagnelie (2012); Shannon (2012).

69. Perlmutter và Mink (2006); Lyons (2011).

70. Koch và cộng sự (2006).

71. Schalk (2008). Xem phần tổng quan những công nghệ mới nhất trong Berger và cộng sự (2008). Xem trường hợp mà điều này có thể dẫn đến tăng cường trí tuệ trong Warwick (2002).

72. Một số ví dụ: Bartels và cộng sự (2008); Simeral và cộng sự (2011); Krusienski và Shih (2011); Pasqualotto và cộng sự (2012).

73. E.g. Hinke và cộng sự (1993).

74. Có một số ngoại lệ cho điều này, đặc biệt là trong xử lý cảm giác ban đầu. Ví dụ, vỏ não thị giác chính sử dụng một kiểu “ánh xạ” võng mạc (retinotopic mapping), nghĩa là các tổ hợp thần kinh liền kề nhận đầu vào từ các vùng võng mạc liền kề (mặc dù các cột thị giác lớn (ocular dominance column) có phần làm phức tạp quá trình ánh xạ).

75. Berger và cộng sự (2012); Hampson và cộng sự (2012).

76. Một số phần cấy ghép não đòi hỏi hai hình thức học tập: thiết bị học tập để giải thích các phép biểu diễn thần kinh của sinh vật và sinh vật học cách sử dụng hệ thống bằng cách tạo ra các mẫu hình phóng xung thần kinh phù hợp (Carmena và cộng sự 2003).

77. Có ý kiến cho rằng chúng ta nên coi các tập thể (tập đoàn, đoàn thể, chính phủ, nhà thờ,...) là tác tử trí tuệ nhân tạo, thực thể có cơ quan thụ cảm và cơ quan chấp hành, có thể trình bày kiến thức và thực hiện suy luận và hành động (ví dụ: Kuipers [2012]; so sánh với Huebner [2008] để xem thêm một cuộc thảo luận về việc liệu các phép biểu diễn tập thể có thể tồn tại). Chúng rõ ràng rất mạnh mẽ và thành công về mặt sinh thái, mặc dù khả năng và trạng thái bên trong của chúng khác với con người.

78. Hanson (1995, 2000); Berg và Rietz (2003).

79. Ví dụ, tại nơi làm việc, chủ lao động có thể sử dụng máy phát hiện nói dối để trấn áp hành vi trộm cắp và trốn tránh của nhân viên, bằng cách hỏi nhân viên vào cuối mỗi ngày làm việc xem có ăn cắp gì không và có làm việc hết sức hay không. Các nhà lãnh đạo chính trị và kinh doanh cũng có thể được hỏi liệu họ có hết lòng theo đuổi lợi ích của các cổ đông hoặc thành viên của họ hay không. Những kẻ độc tài có thể sử dụng chúng để nhắm mục tiêu vào các tướng lĩnh trong chế độ hoặc nghi ngờ những kẻ gây rối trong quần thể rộng lớn hơn.

80. Người ta có thể tưởng tượng các kỹ thuật hình ảnh thần kinh có thể phát hiện các đặc điểm thần kinh của nhận thức có động lực. Nếu không có khả năng phát hiện ra hiện tượng tự nói dối, công nghệ phát hiện nói dối sẽ có lợi cho những cá nhân tin vào sự “tuyên truyền” của bản thân. Các thử nghiệm tốt hơn cho các bài kiểm tra tự nói dối cũng có thể được sử dụng để rèn luyện tính hợp lý và nghiên cứu tính hiệu quả của các biện pháp can thiệp nhằm giảm bớt thiên kiến.

81. Bell và Gemmel (2009). Một ví dụ ban đầu được tìm thấy trong công trình của Deb Roy đến từ MIT, người đã ghi lại từng khoảnh khắc của con trai mình trong ba năm đầu đời. Phân tích dữ liệu nghe nhìn này đang mang lại thông tin về phát triển ngôn ngữ; xem Roy (2012).

82. Tăng trưởng tổng dân số thế giới của loài người sẽ chỉ đóng góp một yếu tố nhỏ. Các kịch bản liên quan đến trí tuệ máy có thể thấy dân số thế giới (bao gồm cả trí tuệ kỹ thuật số) bùng nổ gấp nhiều lần trong một khoảng thời gian ngắn. Nhưng con đường đến

siêu trí tuệ đó liên quan đến trí tuệ nhân tạo hoặc giả lập hoàn chỉnh não bộ, vì vậy chúng ta không cần phải xem xét nó trong tiểu mục này.

83. Vinge (1993).

CHƯƠNG 3: CÁC DẠNG THỨC SIÊU TRÍ TUỆ

1. Vernor Vinge đã sử dụng thuật ngữ “siêu trí tuệ yếu” để chỉ những tâm trí con người được tăng tốc như vậy (Vinge 1993).

2. Ví dụ, nếu một hệ thống rất nhanh có thể làm mọi thứ mà bất kỳ con người nào có thể làm ngoại trừ nhảy điệu mazurka, chúng ta vẫn nên gọi nó là một siêu trí tuệ tốc độ. Mối quan tâm của chúng ta nằm ở những khả năng nhận thức cốt lõi có ý nghĩa kinh tế hoặc chiến lược.

3. Ít nhất một tốc độ tăng gấp triệu lần so với bộ não của con người là có thể, như có thể thấy bằng cách xem xét sự khác biệt về tốc độ và năng lượng của các quá trình não liên quan so với xử lý thông tin hiệu quả hơn. Tốc độ của ánh sáng lớn hơn một triệu lần so với truyền thần kinh, các luồng xung (spike) của synapse tiêu tan nhiệt nhiều hơn một triệu lần so với nhiệt độ cần thiết và tần số bóng bán dẫn hiện tại nhanh hơn một triệu lần so với tần số của neuron phát xung (spiking) (Yudkowsky [2008a]; xem thêm Drexler [1992]). Giới hạn cuối cùng của siêu trí tuệ tốc độ bị giới hạn bởi độ trễ giao tiếp tốc độ ánh sáng, giới hạn lượng tử về tốc độ chuyển trạng thái và thể tích cần thiết để chứa đựng tâm trí (Lloyd 2000). “Laptop tối thượng mà Lloyd (2000) mô tả sẽ chạy một bản giả lập não bộ $1,4 \times 10^{21}$ FLOPS với tốc độ có thể lên đến $3,8 \times 10^{29}$ (nếu bản giả lập có thể được song song hóa phù hợp). Tuy nhiên, kiến trúc của Lloyd không có mục đích công nghệ hợp lý mà chỉ để minh họa những hạn

chế về tính toán có thể dễ dàng suy ra được từ các định luật vật lý cơ bản.

4. Với các bản giả lập, cũng có một vấn đề là một tâm trí giống con người có thể tiếp tục làm việc gì đó trước khi phát điên hoặc rơi vào lối mòn. Ngay cả với sự đa dạng về tác vụ và ngày nghỉ thường xuyên, không có gì chắc chắn rằng một tâm trí giống con người có thể sống hàng ngàn năm chủ quan mà không có các vấn đề tâm lý. Hơn nữa, nếu tổng dung lượng bộ nhớ bị hạn chế, hậu quả của việc có một lượng neuron giới hạn, thì việc học tích lũy không thể tiếp tục vô tận: ngoài một số điểm, tâm trí phải bắt đầu quên đi một điều cho mỗi điều mới mà nó học được. (Trí tuệ nhân tạo có thể được thiết kế để cải thiện những vấn đề tiềm ẩn này.)

5. Theo đó, các cơ chế nano di chuyển với tốc độ khiêm tốn 1 m/s có thang thời gian điển hình là nano giây. Xem phần 2.3.2 của Drexler (1992). Robin Hanson đề cập đến các robot “tinkerbells” 7-mm di chuyển với tốc độ gấp 260 lần tốc độ bình thường (Hanson 1994).

6. Hanson (2012).

7. “Trí tuệ tập thể” không phải là sự song song hóa ở mức độ thấp của phần cứng máy tính mà là sự song song hóa ở cấp độ của các tác tử tự trị thông minh như con người. Việc thực hiện một mô phỏng duy nhất trên một máy song song diện rộng có thể dẫn đến siêu trí tuệ tốc độ nếu máy tính song song đủ nhanh: nó sẽ không tạo ra trí tuệ tập thể.

8. Những cải tiến về tốc độ hoặc chất lượng của các thành phần riêng lẻ cũng có thể ảnh hưởng gián tiếp đến hiệu suất của trí tuệ tập thể, nhưng ở đây chúng tôi chủ yếu đưa ra những cải tiến như vậy

dưới hai hình thức siêu trí tuệ khác trong kiểu phân loại của chúng tôi.

9. Người ta đã lập luận rằng mật độ dân số cao hơn đã gây ra cuộc Cách mạng Hậu kỳ Đồ đá và việc vượt quá ngưỡng tích lũy văn hóa phức tạp nhất định trở nên dễ dàng hơn nhiều (Powell và cộng sự 2009).

10. Internet thì sao? Có vẻ như vẫn chưa đạt được mức tăng siêu lớn. Có lẽ cuối cùng nó sẽ làm như vậy. Phải mất hàng thế kỷ hoặc thiên niên kỷ để các ví dụ khác được liệt kê ở đây tiết lộ tiềm năng đầy đủ của chúng.

11. Rõ ràng thí nghiệm tưởng tượng này không được cố tình đưa ra với tính thực tế. Một hành tinh đủ lớn để duy trì bảy triệu triệu người với công nghệ hiện tại sẽ nổ tung, trừ khi nó được tạo ra từ vật chất rất nhẹ hoặc rỗng và được giữ bằng phương tiện nhân tạo hoặc các phương tiện nhân tạo khác. (Một quả cầu Dyson hoặc một thế giới Shell có thể là một giải pháp tốt hơn.) Lịch sử sẽ khác trên một bề mặt rộng lớn như vậy. Hãy đặt tất cả điều này sang một bên.

12. Trọng tâm của chúng ta ở đây là các thuộc tính chức năng của một trí tuệ thống nhất, không phải là câu hỏi liệu một trí tuệ như vậy sẽ có chất lượng hay liệu nó sẽ là một tâm trí về mặt sở hữu kinh nghiệm có ý thức chủ quan. (Tuy nhiên, người ta có thể suy ngẫm, loại kinh nghiệm có ý thức nào có thể nảy sinh từ dạng trí tuệ tích hợp có tính thống nhất thấp hoặc cao hơn não người. Về một số quan điểm về ý thức, như lý thuyết không gian làm việc toàn cầu, có vẻ như người ta có thể mong đợi bộ não thống nhất hơn có ý thức mạnh mẽ hơn. So sánh với Baars [1997], Shanahan [2010], và Schwitzgebel [2013].)

13. Ngay cả những nhóm nhỏ người bị cô lập trong một thời gian vẫn có thể được hưởng lợi từ kết quả trí tuệ của trí tuệ tập thể lớn hơn. Ví dụ, ngôn ngữ họ sử dụng có thể được phát triển bởi một cộng đồng ngôn ngữ lớn hơn nhiều và các công cụ họ sử dụng có thể đã được phát minh trong một dân số lớn hơn nhiều trước khi nhóm nhỏ bị cô lập. Nhưng ngay cả khi một nhóm nhỏ luôn bị cô lập, nó vẫn có thể là một phần của trí tuệ tập thể lớn hơn về ngoài, cụ thể là trí tuệ tập thể bao gồm không chỉ hiện tại mà tất cả các thế hệ tổ tiên như chào đón, một tập hợp có thể hoạt động như một hệ thống xử lý thông tin đường thẳng (feed-forward).

14. Theo luận đề Church Turing, tất cả các hàm tính toán đều có thể tính toán được bằng máy Turing. Vì bất kỳ một trong ba hình thức siêu trí tuệ nào cũng có thể mô phỏng một máy Turing (nếu được cấp quyền truy cập vào bộ nhớ không giới hạn và được phép hoạt động vô thời hạn), chúng là tương đương về mặt tính toán theo tiêu chí này. Thật vậy, một người bình thường (được cung cấp nguồn giấy không giới hạn và thời gian không giới hạn) cũng có thể làm một máy Turing, và do đó cũng là tương đương theo tiêu chí này. Tuy nhiên, điều quan trọng đối với mục đích của chúng ta là những gì các hệ thống khác nhau này có thể đạt được trong thực tế, với bộ nhớ hữu hạn và trong thời gian hợp lý. Và sự đa dạng về hiệu quả lớn đến mức người ta có thể dễ dàng phân biệt. Ví dụ, một cá nhân điển hình có chỉ số IQ 85 có thể được dạy để làm máy Turing. (Thậm chí có thể huấn luyện một số con tinh tinh có năng khiếu đặc biệt và ngoan ngoãn để làm điều này.) Tuy nhiên, đối với tất cả các ý định và mục đích thực tế, một cá nhân như vậy có lẽ không có khả năng, chẳng

hạn, phát triển độc lập thuyết tương đối rộng hoặc giành được giải thưởng Field.

15. Lối truyền miệng có thể tạo ra các tác phẩm văn học tuyệt vời (như sử thi Homer) nhưng có lẽ một số tác giả cũng sở hữu tài năng khác thường.

16. Trừ khi nó chứa các thành phần nội tại có siêu trí tuệ tốc độ hoặc chất lượng.

17. Chúng tôi không có khả năng xác định tất cả những vấn đề này có thể một phần là do thiếu cố gắng: có rất ít thời gian được dành để chi tiết hóa các công việc trí tuệ mà không một cá nhân và tổ chức hiện tại nào có thể thực hiện được. Nhưng cũng có thể là ngay cả việc khái niệm hóa một số công việc này cũng là một trong những công việc mà chúng ta hiện đang chưa có đủ não bộ để thực hiện

18. So sánh với Boswell (1917); xem thêm Walker (2002).

19. Điều này chủ yếu xảy ra trong các đợt bùng nổ ngắn trong một tập hợp các neuron, hầu hết có tốc độ bắn chậm hơn. (Gray và McCormick 1996; Steriade và cộng sự 1998). Có một số neuron (các “neuron rung đập”, còn được gọi là các tế bào “bùng nổ theo nhịp độ nhanh”) có thể đạt tới tần số kích hoạt cao tới 750 Hz, nhưng chúng có vẻ là những ngoại lệ.

20. Feldman và Ballard (1982).

21. Tốc độ dẫn truyền phụ thuộc vào đường kính sợi trục (sợi trục dày hơn nhanh hơn) và vào việc liệu sợi trục có bị myelin hóa hay không. Trong hệ thống thần kinh trung ương, độ trễ truyền có thể dao động từ dưới một phần nghìn giây đến tối đa 100 ms (Kandel và cộng sự 2000). Truyền trong sợi quang là khoảng 68% vận tốc ánh sáng (vì

chỉ số khúc xạ của vật liệu). Các điện có tốc độ gần bằng nhau, 59-77% vận tốc ánh sáng.

22. Điều này giả sử vận tốc tín hiệu là 70% vận tốc ánh sáng. Giả sử 100% vận tốc ánh sáng sẽ tăng ước tính lên $1,8 \times 10^{18}$ m³.

23. Số lượng tế bào thần kinh trong não người trưởng thành được ước tính là $86,1 \pm 8,1$ tỷ, một con số được tạo ra bằng cách hòa tan não và phân tách nhân tế bào, đếm những hạt được nhuộm bằng một chất chỉ thị riêng cho neuron. Trước đây, các ước tính xung quanh 75-125 tỷ neuron là phổ biến. Điều này dựa trên việc đếm mật độ tế bào thủ công ở các vùng nhỏ đại diện (Azevedo và cộng sự 2009).

24. Whitehead (2003).

25. Các hệ thống xử lý thông tin rất có thể sử dụng các quy trình quy mô phân tử để tính toán và lưu trữ dữ liệu và đạt được ít nhất kích thước hành tinh trong phạm vi. Tuy nhiên, các giới hạn vật lý cuối cùng đối với tính toán được thiết lập bởi cơ học lượng tử, thuyết tương đối rộng và nhiệt động lực học là vượt xa cấp độ siêu máy tính này (Sandberg 1999; Lloyd 2000).

26. Stansberry và Kudritzki (2012). Điện được sử dụng trong các trung tâm dữ liệu trên toàn thế giới lên tới 1,1-1,5% tổng lượng điện sử dụng (Koomey 2011). Xem thêm Muehlhauser và Salamon (2012).

27. Đây là một sự đơn giản hóa quá mức. Số lượng các khối mà bộ nhớ làm việc có thể duy trì phụ thuộc vào cả thông tin và tác vụ; tuy nhiên, rõ ràng nó bị giới hạn ở một số lượng khối nhỏ. Xem Miller (1956) và Cowan (2001).

28. Có thể lấy ví dụ: khó khăn trong việc học các khái niệm Boole (các phạm trù được xác định bởi các quy tắc logic) tỷ lệ thuận với độ

dài của công thức mệnh đề tương đương logic ngắn nhất. Thông thường, ngay cả các công thức chỉ dài ba đến bốn chữ cũng rất khó học. Xem Feldman (2000).

29. Xem Landauer (1986). Nghiên cứu này dựa trên các ước tính thực nghiệm về tỷ lệ học và quên ở người. Có tính đến việc học ngầm có thể đẩy ước tính lên một chút. Nếu giả sử dung lượng lưu trữ xấp xỉ 1 bit cho mỗi synapse, chúng ta sẽ có giới hạn trên về dung lượng bộ nhớ của con người vào khoảng 10¹⁵ bit. Để biết tổng quan về các ước tính khác nhau, xem Phụ lục A của Sandberg và Bostrom (2008).

30. Nhiều kênh có thể kích hoạt tiềm năng hành động, và nhiều synapse tạo ra sự thay đổi đáng kể về cường độ của tín hiệu truyền đi. Các hệ thống thần kinh dường như đã phát triển để tạo ra nhiều sự đánh đổi giữa khả năng chịu nhiễu và chi phí (khối lượng, kích thước, độ trễ thời gian); xem Faisal và cộng sự (2008). Ví dụ, các sợi trục không thể mỏng hơn 0,1 μm nếu không việc mở ngẫu nhiên các kênh ion sẽ tạo ra tiềm năng hành động tự phát (Faisal và cộng sự 2005).

31. Trachtenberg và cộng sự (2002).

32. Về mặt bộ nhớ và công suất tính toán, mặc dù không phải là về hiệu quả năng lượng. Máy tính nhanh nhất thế giới tại thời điểm viết cuốn sách này là “Thiên Hà-2” của Trung Quốc, đã thay thế Cray Inc. Titan vào tháng 6 năm 2013 với hiệu suất 33,86 petaFLOPS. Nó sử dụng 17,6 MW năng lượng, gấp gần một triệu lần con số ~20 W của não.

33. Lưu ý rằng khảo sát về các nguồn lợi thế của trí tuệ máy này là không có liên kết: lập luận của chúng tôi là đúng ngay cả khi một

số mục được liệt kê là ảo, miễn là có ít nhất một nguồn có thể cung cấp lợi thế đủ lớn.

CHƯƠNG 4: ĐỘNG LỰC CỦA SỰ BÙNG NỔ TRÍ TUỆ

1. Hệ thống có thể không đạt được một trong những đường cơ sở này tại bất kỳ điểm được xác định rõ ràng nào. Thay vào đó, có thể có một khoảng thời gian mà trong đó, hệ thống dần dần có thể vượt trội hơn nhóm nghiên cứu bên ngoài về số lượng ngày càng tăng của các tác vụ phát triển cải thiện hệ thống.

2. Trong nửa thế kỷ qua, ít nhất một kịch bản đã được công nhận rộng rãi, trong đó trật tự thế giới hiện tại sẽ chấm dứt trong vài phút hoặc vài giờ: chiến tranh nhiệt hạch toàn cầu.

3. Điều này sẽ phù hợp với quan sát rằng hiệu ứng Flynn (sự gia tăng liên tục về chỉ số IQ đo được trong hầu hết các cộng đồng với tốc độ khoảng

3 điểm IQ mỗi thập kỷ trong 60 năm qua hoặc tương đương) dường như đã dừng lại hoặc thậm chí đảo ngược trong những năm gần đây ở một số nước phát triển mạnh như Vương quốc Anh, Đan Mạch và Na Uy (Teasdale và Owen 2008; Sundet và cộng sự 2004). Nguyên nhân gây ra hiệu ứng Flynn trong quá khứ (và việc liệu nó có đại diện cho bất kỳ mức tăng nào trong trí tuệ tổng quát hay kỹ năng chỉ đơn thuần được cải thiện trong việc giải các câu đố kiểu IQ) đã là chủ đề được tranh luận rộng rãi và vẫn chưa có lời giải. Ngay cả khi hiệu ứng Flynn (ít nhất là một phần) phản ánh mức tăng nhận thức thực sự và khi nó đang suy giảm dần hoặc thậm chí đảo ngược, điều này không chứng minh rằng chúng ta đã đạt được hiệu suất giảm dần trong bất kỳ nguyên nhân cơ bản nào chịu trách nhiệm cho hiệu ứng Flynn được quan sát trong quá khứ. Thay vào đó, sự suy giảm

hoặc đảo ngược có thể là do một số yếu tố bất lợi riêng rẽ mà trong trường hợp khác có thể tạo ra sự suy giảm thậm chí còn lớn hơn.

4. Bostrom và Roache (2011).

5. Liệu pháp gen sinh dưỡng có thể loại bỏ độ trễ trưởng thành, nhưng về mặt kỹ thuật, biện pháp này khó thực hiện hơn nhiều so với các biện pháp can thiệp dòng mầm và cũng có tiềm năng cuối cùng thấp hơn.

6. Tăng trưởng sản lượng kinh tế toàn cầu trung bình mỗi năm trong giai đoạn 1960-2000 là 4,3% (Isaksson 2007). Chỉ một phần của sự tăng trưởng năng suất này là do sự gia tăng về hiệu quả tổ chức. Một số mạng lưới hoặc quy trình tổ chức cụ thể tất nhiên đang cải thiện với tốc độ nhanh hơn nhiều.

7. Nhiều ràng buộc và sự đánh đổi mà não bộ sinh học đã trải qua khi tiến hóa sẽ được giảm nhẹ khi chuyển sang môi trường kỹ thuật số. Ví dụ, kích thước não bị giới hạn bởi kích thước đầu và một cái đầu quá lớn sẽ gây khó khăn cho sinh nở. Một bộ não lớn cũng tiêu tốn thêm dưỡng chất và là “gánh nặng” cản trở việc di chuyển. Khả năng kết nối giữa các vùng não nhất định có thể bị hạn chế bởi các hạn chế về không gian: khối lượng chất trắng lớn hơn đáng kể so với thể tích của chất xám mà nó liên kết. Quá trình “tản nhiệt” bị giới hạn bởi lưu lượng máu và có thể gần với giới hạn trên cho một hiệu suất làm việc chấp nhận được. Hơn nữa, các tế bào thần kinh sinh học ồn ào, chậm chạp và cần được bảo vệ, bảo trì và tiếp tế liên tục bởi các tế bào thần kinh đệm và mạch máu (góp phần vào sự chật chội bên trong hộp sọ). Xem Bostrom và Sandberg (2009b).

8. Yudkowsky (2008a, 326). Xem phần thảo luận cập nhật hơn trong Yudkowsky (2013).

9. Để đơn giản hóa bản vẽ, chúng tôi chỉ minh họa khả năng nhận thức như một tham số một chiều. Điều này không quá quan trọng đối với vấn đề đang thảo luận. Thay vào đó, người ta có thể biểu diễn một khả năng nhận thức như một siêu diện trong một không gian đa chiều.

10. Lin và cộng sự (2012).

11. Ta có thể có được sự gia tăng nhất định về trí tuệ tập thể chỉ đơn giản bằng cách tăng số lượng các thành phần của nó. Làm vậy ít nhất sẽ cho phép hiệu suất tổng thể tốt hơn trên các tác vụ có thể dễ dàng thực hiện song song. Tuy nhiên, để gặt hái toàn bộ lợi ích từ một sự “bùng nổ dân số” như vậy, người ta cũng cần phải đạt được sự phối hợp ở một mức độ nhất định (hơn mức tối thiểu) giữa các thành phần.

12. Sự khác biệt giữa tốc độ và chất lượng của trí tuệ dù sao cũng khá mờ nhạt trong trường hợp các hệ thống AI phi hình thái thần kinh (non- neuromorphic).

13. Rajab và cộng sự (2006, 41–52).

14. Người ta đã đề xuất rằng sử dụng các mạch tích hợp có thể cấu hình (FPGA) thay cho các bộ xử lý đa năng có thể tăng tốc độ tính toán trong các bản mô phỏng mạng lưới thần kinh lên tới hai cấp khuếch đại (Markram 2006). Một nghiên cứu về lập mô hình khí hậu độ phân giải cao theo petaFLOP cho thấy giảm 24 đến 34 lần chi phí và khoảng hai cấp khuếch đại về yêu cầu năng lượng bằng cách sử dụng một phiên bản chip xử lý nhúng đã tùy chỉnh (Wehner và cộng sự 2008).

15. Nordhaus (2007). Có rất nhiều bài tổng quan về ý nghĩa khác nhau của định luật Moore; ví dụ, xem Tuomi (2002) và Mack (2011).

16. Nếu sự phát triển diễn ra đủ chậm, dự án có thể tự tận dụng tiến bộ đang được thực hiện trong giai đoạn chuyển tiếp bởi thế giới bên ngoài, chẳng hạn như những tiến bộ trong khoa học máy tính được các nhà nghiên cứu ở trường đại học thực hiện và những cải tiến về phần cứng của ngành công nghiệp bán dẫn.

17. “Mái hiên thuật toán” có lẽ ít khả năng hơn xảy ra, nhưng sẽ có một ngoại lệ khi chúng ta có sẵn loại phần cứng kỳ lạ như điện toán lượng tử để chạy các thuật toán mà trước đây không thể chạy được. Người ta cũng có thể lập luận rằng các mạng lưới thần kinh và học máy sâu là các “mái hiên thuật toán”: vì quá tốn kém để có thể hoạt động tốt khi lần đầu tiên được phát minh, chúng bị gác lại trong một thời gian, sau đó sẽ được dùng lại khi có các bộ xử lý đồ họa nhanh khiến việc chạy chúng dễ dàng hơn. Bây giờ chúng là kẻ chiến thắng.

18. Othế giới là một phần của năng lực tối ưu hóa của thế giới, được áp dụng để cải thiện hệ thống đang xem xét. Đối với một dự án hoạt động trong sự cô lập hoàn toàn, và không nhận được sự hỗ trợ liên tục từ thế giới bên ngoài, chúng tôi có Othế giới ≈ 0 , mặc dù dự án phải bắt đầu bằng một nguồn tài nguyên (máy tính, các khái niệm khoa học, sự có đào tạo,...) có nguồn gốc từ toàn bộ nền kinh tế thế giới và nhiều thế kỷ phát triển.

19. Điều có liên quan nhất đến khả năng nhận thức của AI hạt giống ở đây là khả năng thiết kế thông minh để cải thiện bản thân, tức là khả năng tự khuếch đại trí tuệ. (Nếu AI hạt giống giỏi tăng cường hệ thống khác, mà các hệ thống này lại giỏi việc tăng cường cho AI hạt giống, thì chúng ta có thể xem chúng là hệ thống con của một hệ thống lớn hơn và tập trung phân tích hệ thống tổng thể này.)

20. Điều này giả định rằng sức chống đối không được biết đến như một giá trị quá cao đến mức không khuyến khích đầu tư hoặc chuyển hướng đầu tư sang một dự án thay thế nào đó.

21. Một ví dụ tương tự được thảo luận trong Yudkowsky (2008b).

22. Vì đầu vào đã tăng (ví dụ: số tiền đầu tư xây dựng các xưởng mới và số người làm việc trong ngành công nghiệp bán dẫn) nên bản thân định luật Moore đã không đưa ra một sự tăng trưởng nhanh như vậy nếu chúng ta kiểm soát mức tăng đầu vào này. Tuy nhiên, kết hợp với những tiến bộ trong phần mềm, thời gian nhân đôi hiệu suất trên mỗi đơn vị đầu vào là 18 tháng có thể sẽ hợp lý hơn về mặt lịch sử.

23. Và ngay cả khi sự tiến bộ về phía đường cơ sở con người bị chậm.

24. Một số nỗ lực dự kiến đã được thực hiện để phát triển ý tưởng về một sự bùng nổ trí tuệ trong khuôn khổ lý thuyết tăng trưởng kinh tế; ví dụ, xem Hanson (1998b); Jones (2009); Salamon (2009). Những nghiên cứu này đã chỉ ra tiềm năng tăng trưởng cực kỳ nhanh chóng do sự xuất hiện của trí tuệ kỹ thuật số, nhưng vì lý thuyết tăng trưởng nội sinh tương đối kém phát triển ngay cả đối với các ứng dụng trong quá khứ và đương đại, bất kỳ ứng dụng nào cho bối cảnh tương lai đứt gãy trong giai đoạn này đều nên được coi như một nguồn cung cấp các khái niệm và cân nhắc hữu ích chứ không phải một tình huống thực tiễn có khả năng đưa ra các dự báo có căn cứ. Để biết tổng quan các nỗ lực mô hình hóa về mặt toán học một điểm kỳ dị công nghệ, xem Sandberg (2010).

25. Hiển nhiên cũng có thể sẽ không có quá trình cất cánh nào cả. Nhưng vì, như đã lập luận trước đó, siêu trí tuệ xem ra khả thi về mặt

kỹ thuật, nên việc không xảy ra quá trình cất cánh có thể là do sự can thiệp của một số yếu tố tàn phá, chẳng hạn như một thảm họa diệt vong. Nếu siêu trí tuệ mạnh xuất hiện không phải ở dạng trí tuệ nhân tạo hay giả lập tổng thể não bộ mà thông qua một trong những con đường khác mà chúng ta đã xem xét ở trên, thì một quá trình cất cánh chậm hơn sẽ có nhiều khả năng xảy ra.

CHƯƠNG 5: LỢI THẾ CẠNH TRANH QUYẾT ĐỊNH

1. Một tư duy dạng phần mềm có thể chạy trên một máy duy nhất trái ngược với một mạng máy tính trên toàn thế giới; nhưng đây không phải khái niệm “tập trung” của chúng tôi. Thay vào đó, điều chúng tôi quan tâm ở đây là mức độ mà năng lực, đặc biệt là năng lực có được từ khả năng công nghệ, sẽ được tập trung trong các giai đoạn tiến bộ cao của, hoặc ngay sau, cuộc cách mạng trí tuệ máy.

2. Chẳng hạn, sự truyền bá công nghệ của các sản phẩm tiêu dùng có xu hướng chậm hơn ở các nước đang phát triển (Talukdar và cộng sự 2002). Xem thêm Keller (2004) và Ngân hàng Thế giới (2008).

3. Các tài liệu kinh tế liên quan đến lý thuyết về doanh nghiệp có liên quan với tư cách là một điểm đối chiếu cho nội dung thảo luận. Trong số đó, Coase (1937) có một vị thế đặc biệt. Ví dụ, xem thêm Canbäck và cộng sự (2006); Milgrom và Roberts (1990); Hart (2008); Simester và Knez (2002).

4. Mặt khác, việc đánh cắp một AI hạt giống là đặc biệt dễ, vì nó bao gồm phần mềm có thể được truyền bằng phương pháp điện tử hoặc lưu trên thiết bị nhớ di động.

5. Nếu chúng ta mô hình hóa tình huống này sao cho trong đó, thời gian trễ giữa các dự án được rút ra từ một sự phân phối bình

thường, thì khoảng cách có thể có giữa dự án dẫn đầu và dự án bám đuôi gần nhất cũng sẽ phụ thuộc vào số lượng dự án. Nếu có một số lượng lớn các dự án, khoảng cách giữa hai dự án dẫn đầu có thể sẽ nhỏ ngay cả khi phương sai của phân phối là không quá cao (mặc dù khoảng cách dự kiến giữa dự án dẫn đầu và dự án thứ hai giảm rất chậm với số lượng đối thủ cạnh tranh nếu thời gian hoàn thành được phân phối bình thường). Tuy nhiên, khả năng không thể có một số lượng lớn các dự án được cung cấp đủ tài nguyên để trở thành đối thủ nặng ký. (Có thể có nhiều dự án hơn nếu có một số lượng lớn các phương pháp tiếp cận cơ bản khác nhau để theo đuổi, nhưng trong trường hợp này, nhiều cách tiếp cận trong số đó có khả năng đi vào ngõ cụt.) Như đã đề xuất, theo kinh nghiệm, chúng ta dường như thấy rằng thông thường không có nhiều hơn một số đối thủ cạnh tranh nghiêm túc theo đuổi bất kỳ mục tiêu công nghệ cụ thể nào. Tình hình sẽ hơi khác trong một thị trường tiêu dùng có nhiều “ngách” cho các sản phẩm chỉ khác nhau đôi chút và rào cản gia nhập là thấp. Có rất nhiều dự án thiết kế áo phông chỉ có một người tiến hành, nhưng chỉ có một số ít công ty trên thế giới phát triển các thế hệ card đồ họa mới. (Hai công ty AMD và NVIDIA hiện đang tận hưởng vị thế gần như độc quyền, mặc dù Intel cũng đang cạnh tranh ở phân khúc có hiệu năng thấp hơn của thị trường.)

6. Barber (1991) cho rằng nền văn hóa Ngưỡng Thiều (năm 5000-3000 TCN) có thể đã sử dụng lụa. Sun và cộng sự (2012), dựa trên các nghiên cứu di truyền, ước tính rằng con người đã biết nuôi tằm lấy tơ từ khoảng 4.100 năm trước.

7. Cook (1984, 144). Câu chuyện này có thể không chống lại được sự soi mói trong lịch sử, giống như câu chuyện của Procopius (Wars

VIII.xvii.1-7) về việc những con tằm được những tu sĩ lang thang, ẩn mình trong những hang tre rỗng đưa đến Byzantium ra sao (Hunt 2011).

8. Wood (2007); Temple (1986).

9. Các nền văn hóa tiền Columbus đã có bánh xe nhưng chỉ sử dụng nó cho đồ chơi (có lẽ do thiếu thú kéo xe tốt).

10. Koubi (1999); Lerner (1997); Koubi và Lalman (2007); Zeira (2011); Judd và cộng sự (2012).

11. Ước tính từ nhiều nguồn khác nhau. Khoảng cách thời gian thường có phần tùy ý, tùy thuộc vào cách định nghĩa về năng lực “tương đương”. Radar đã được sử dụng bởi ít nhất hai quốc gia trong vòng vài năm kể từ khi được giới thiệu, nhưng con số chính xác trong nhiều tháng rất khó để có được.

12. Ellis (1999).

13. RDS-6 năm 1953 là thử nghiệm đầu tiên của một quả bom có phản ứng nhiệt hạch, nhưng RDS-37 năm 1955 là quả bom nhiệt hạch đầu tiên đúng nghĩa, trong đó đa phần sức mạnh đến từ phản ứng nhiệt hạch.

14. Chưa được xác nhận.

15. Các thử nghiệm tiến hành năm 1989, dự án bị hủy bỏ năm 1994.

16. Hệ thống đã triển khai, có tầm bắn hơn 5.000 km.

17. Tên lửa Polaris mua từ Mỹ.

18. Công việc hiện tại đang được tiến hành trên tên lửa Taimur, có khả năng dựa trên tên lửa của Trung Quốc.

19. Tên lửa RSA-3 được thử nghiệm vào năm 1989-1990 được dùng để phóng vệ tinh và/hoặc như một tên lửa đạn đạo liên lục địa

(ICBM).

20. MIRV = tên lửa hồi quyển đa đầu đạn định hướng độc lập, công nghệ cho phép một tên lửa đạn đạo duy nhất mang được nhiều đầu đạn có thể được lập trình để bắn trúng các mục tiêu khác nhau.

21. Hệ thống Agni V chưa hoạt động.

22. Bostrom (2006c). Người ta có thể tưởng tượng một thể đơn nhất tồn tại vô hình (ví dụ: một siêu trí tuệ với công nghệ tiên tiến hoặc hiểu biết sâu sắc đến mức có thể điều khiển một cách khéo léo các sự kiện thế giới mà không cần bất kỳ sự can thiệp nào của con người); hoặc một thể đơn nhất tự nguyện áp đặt những hạn chế rất nghiêm ngặt đối với việc thực thi quyền lực của chính mình (ví dụ, tự giới hạn bản thân để đảm bảo rằng các quy tắc quốc tế nhất định theo quy định của Hiệp ước hay các nguyên tắc tự do được tôn trọng). Thể đơn nhất cụ thể bất kỳ có khả năng xuất hiện cao hay thấp tất nhiên là một câu hỏi thực tiễn; nhưng về mặt khái niệm, ít nhất, có thể có một thể đơn nhất tốt hay không tốt, một thể đơn nhất biến hóa hay ít thay đổi, một thể độc tôn có xu hướng áp bức kìm kẹp hay giống với một quy luật tự nhiên bổ sung hơn là một tên bạo ngược.

23. Jones (1985, 344).

24. Việc Dự án Manhattan được thực hiện trong thời chiến có thể có ý nghĩa đáng kể. Nhiều nhà khoa học tham gia đã tuyên bố rằng mình chủ yếu được thúc đẩy bởi tình hình thời chiến và nỗi sợ rằng Đức Quốc xã có thể phát triển vũ khí nguyên tử trước quân Đồng minh. Nhiều chính phủ có thể gặp khó khăn trong việc huy động một nỗ lực chuyên sâu và bí mật tương tự trong thời bình. Chương trình Apollo, một siêu dự án khoa học/ kỹ thuật mang tính biểu tượng

khác, đã có động lực mạnh mẽ từ sự cạnh tranh trong Chiến tranh Lạnh.

25. Mặc dù ngay cả khi họ đang tìm kiếm cật lực, việc họ tỏ ra như vậy (trước công chúng) vẫn không rõ ràng.

26. Kỹ thuật mã hóa có thể cho phép nhóm cộng tác được phân tán vật lý.

Liên kết yếu duy nhất trong chuỗi giao tiếp có thể là giai đoạn đầu vào, trong đó hành động vật lý của việc gõ phím có thể được quan sát. Nhưng nếu giám sát trong nhà trở nên phổ biến (bằng các thiết bị ghi vi mô), những người quan tâm đến việc bảo vệ quyền riêng tư có thể phát triển các biện pháp đối phó (ví dụ: các khoang đặc biệt có thể “bịt tai” các thiết bị nghe lén). Trong khi không gian vật lý có thể trở nên trong suốt trong thời đại giám sát sắp tới, không gian mạng có thể được bảo vệ nhiều hơn thông qua việc áp dụng rộng rãi các giao thức mật mã mạnh hơn.

27. Một nhà nước toàn trị có thể đòi hỏi các biện pháp cưỡng chế hơn nữa.

Các nhà khoa học trong các lĩnh vực có liên quan có thể bị tập hợp lại và đưa vào trại tập trung.

28. Khi mức độ quan tâm của công chúng là tương đối thấp, một số nhà nghiên cứu có thể “dọa dẫm” công chúng đôi chút vì điều này thu hút sự chú ý vào công trình của họ và làm cho ngành nghiên cứu đó có vẻ quan trọng và thú vị. Khi mức độ quan tâm trở nên lớn hơn, các cộng đồng nghiên cứu có liên quan có thể đổi giọng điệu khi bắt đầu lo lắng về việc cắt giảm tài trợ, quy định và phản ứng dữ dội từ công chúng. Các nhà nghiên cứu trong các ngành lân cận, như khoa học máy tính và robot không liên quan nhiều đến trí tuệ nhân tạo tổng

thể, có thể phản nộ với sự chuyển hướng tài trợ và việc lĩnh vực nghiên cứu của mình mất đi sự quan tâm. Các nhà nghiên cứu này cũng có thể quan sát chính xác rằng công việc của họ không có nguy cơ gây ra một sự bùng nổ trí tuệ nguy hiểm. (Một số tương đồng lịch sử có thể được rút ra với sự nghiệp của ý tưởng về công nghệ nano; xem Drexler [2013].)

29. Đã có thành công ở chỗ họ đã đạt được ít nhất một số điều mà mình đặt ra. Làm thế nào thành công ấy có ý nghĩa rộng hơn (có tính đến hiệu quả về chi phí và nhiều thứ khác) là một việc khó xác định. Ví dụ, trong trường hợp của Trạm Vũ trụ Quốc tế, việc chi phí leo thang và các khoảng trì hoãn đã xuất hiện. Để biết chi tiết về các vấn đề mà dự án gặp phải, xem NASA (2013). Dự án Large Hadron Collider (Máy Gia tốc Hạt Hadron Lớn) đã có một số thất bại đáng kể, nhưng điều này có thể là do sự khác biệt vốn có của nhiệm vụ. Dự án bộ gen người đã đạt được thành công, nhưng dường như đã nhận được sự thúc đẩy từ việc bị buộc phải cạnh tranh với nỗ lực từ công ty tiên phong của Craig Venter. Các dự án quốc tế được tài trợ để nghiên cứu năng lượng nhiệt hạch có kiểm soát đã không đạt được kỳ vọng mặc dù được đầu tư lớn; nhưng một lần nữa, điều này có thể được quy cho việc nhiệm vụ hóa ra khó khăn hơn dự đoán.

30. Quốc hội Mỹ, Phòng Đánh giá Công nghệ (1995).

31. Hoffman (2009); Rhodes (2008).

32. Rhodes (1986).

33. Tổ chức giải mã của Hải quân Mỹ, OP-20-G, dường như đã bỏ qua lời mời để có được kiến thức đầy đủ về các phương pháp giải mã Enigma của Anh và không thể thông báo cho các nhà ra quyết định cấp cao hơn của Mỹ về việc Anh chia sẻ bí mật về mật mã

của mình (Burke 2001). Điều này mang lại cho các nhà lãnh đạo Mỹ ấn tượng rằng Anh đang giữ kín thông tin quan trọng, một nguyên nhân gây ra xích mích trong suốt cuộc chiến. Anh đã chia sẻ với chính phủ Liên Xô một số thông tin tình báo mà họ đã lượm lặt được từ các thông tin liên lạc được giải mã của Đức. Cụ thể, họ đã được cảnh báo về sự chuẩn bị của Đức cho Chiến dịch Barbarossa. Nhưng Stalin từ chối tin vào lời cảnh báo, một phần vì người Anh không tiết lộ họ đã lấy được thông tin ra sao.

34. Trong một vài năm, Russell dường như đã ủng hộ lời đe dọa chiến tranh hạt nhân để thuyết phục Nga chấp nhận kế hoạch của Baruch; sau đó, ông là người ủng hộ mạnh mẽ việc giải trừ hạt nhân lẫn nhau (Russell và Griffin 2001). Người ta nói John von Neumann đã tin rằng một cuộc chiến giữa Mỹ và Nga là không thể tránh khỏi, và đã nói: “Nếu bạn nói rằng tại sao không ném bom họ [người Nga] vào ngày mai, tôi sẽ nói tại sao không ném bom ngay hôm nay? Nếu bạn nói hôm nay lúc năm giờ, tôi nói tại sao không phải là một giờ?” Xem Blair [1957, 96].)

35. Baratta (2004).

36. Nếu AI được điều khiển bởi một nhóm người, vấn đề có thể áp dụng cho nhóm người này, mặc dù có thể sẽ có các cách thức mới để đưa ra cam kết đáng tin cậy về một thỏa thuận vào thời điểm này, mà trong trường hợp đó, ngay cả các nhóm người cũng có thể tránh được vấn đề này (về việc lộ bí mật nội bộ và đảo chính đến từ một liên minh nhỏ).

CHƯƠNG 6: SIÊU NĂNG LỰC NHẬN THỨC

1. Theo nghĩa nào thì loài người là một loài thống trị trên Trái đất? Về mặt sinh thái, con người là loài động vật lớn (~50 kg) phổ biến

nhất, nhưng tổng sinh khối khô của toàn nhân loại (~100 tỷ kg) là không quá ấn tượng so với loài kiến, họ Formicidae (300 tỷ-3.000 tỷ kg). Con người, cây trồng và vật nuôi tạo thành một phần rất nhỏ (<0,001) trong tổng sinh khối toàn cầu. Tuy nhiên, đất trồng trọt và đồng cỏ hiện là một trong những hệ sinh thái lớn nhất trên hành tinh, chiếm khoảng 35% diện tích mặt đất không có băng (Foley và cộng sự 2007). Và chúng ta chiếm gần một phần tư năng suất sản xuất chính theo một đánh giá điển hình (Haberl và cộng sự 2007), mặc dù ước tính dao động từ 3% đến hơn 50% tùy thuộc chủ yếu vào các định nghĩa khác nhau của các thuật ngữ liên quan (Haberl và cộng sự 2013). Con người cũng có phạm vi bao phủ địa lý lớn nhất trong số các loài và đứng đầu trong nhiều chuỗi thức ăn nhất.

2. Zalasiewicz và cộng sự (2008).

3. Xem ghi chú đầu tiên của chương này.

4. Nói đúng ra, điều này có thể không hoàn toàn chính xác. Trí tuệ trong loài người có thể là mọi giá trị xuống tận xấp xỉ 0 (ví dụ như phôi thai hoặc bệnh nhân ở trạng thái thực vật vĩnh viễn). Do đó, về mặt định tính, sự khác biệt tối đa về khả năng nhận thức của loài người có lẽ lớn hơn sự khác biệt giữa bất kỳ con người nào và siêu trí tuệ. Nhưng luận điểm ở đây sẽ đứng vững nếu ta chỉ nói về “người trưởng thành bình thường” khi dùng thuật ngữ “con người”.

5. Gottfredson (2002). Xem thêm Carroll (1993) và Deary (2001).

6. Xem Legg (2008). Legg về sơ bộ đã đề xuất đánh giá một tác tử học tăng cường như hiệu suất mong đợi của nó trong tất cả các môi trường có thể đạt được phần thưởng, trong đó mỗi môi trường như vậy sẽ có trọng số được xác định bởi độ phức tạp Kolmogorov của

nó. Chúng tôi sẽ giải thích thêm về học tăng cường trong Chương 12. Xem thêm Dowe và Hernández-Orallo (2012) và Hibbard (2011).

7. Liên quan đến nghiên cứu công nghệ trong các lĩnh vực như công nghệ sinh học và công nghệ nano, điều mà một siêu trí tuệ sẽ nổi trội là thiết kế và mô hình hóa các cấu trúc mới. Chừng nào sự khéo léo trong thiết kế và mô hình hóa không thể thay thế cho quá trình trải nghiệm vật lý, lợi thế về hiệu suất của siêu trí tuệ có thể là phù hợp vì nó khả năng tiếp cận với thiết bị thử nghiệm cần thiết.

8. Ví dụ: Drexler (1992, 2013).

9. Một AI đặc thù lĩnh vực tất nhiên có thể có các ứng dụng thương mại quan trọng, nhưng điều này không có nghĩa là nó sẽ có siêu năng lực năng suất kinh tế. Ví dụ, ngay cả khi một AI đặc thù lĩnh vực kiếm được cho chủ sở hữu của nó vài tỷ đô la mỗi năm, thì vẫn ít hơn so với phần còn lại của nền kinh tế thế giới bốn cấp khuếch đại. Để hệ thống có khả năng gia tăng trực tiếp và đáng kể sản lượng của thế giới, một AI sẽ cần có khả năng thực hiện nhiều loại công việc; tức là, nó sẽ cần năng lực trong nhiều lĩnh vực.

10. Tiêu chí này không loại trừ tất cả các kịch bản mà trong đó AI thất bại. Ví dụ, AI có thể chơi một canh bạc có khả năng thất bại cao. Tuy nhiên, trong trường hợp này, tiêu chí có thể ở dạng (a) AI sẽ đưa ra ước tính không thiên vị về cơ hội thành công thấp của việc đánh bạc và (b) sẽ không có canh bạc nào tốt hơn cho AI mà chúng ta ngày nay có thể nghĩ ra nhưng AI lại bỏ qua.

11. So sánh với Freitas (2000) và Vassar và Freitas (2006).

12. Yudkowsky (2008a).

13. Freitas (1980); Freitas và Merkle (2004, Chương 3); Armstrong và Sandberg (2013).

14. Ví dụ, xem Huffman and Pless (2003), Knill và cộng sự (2000), Drexler (1986).

15. Ước tính này dựa trên ước tính của WMAP (Tàu thăm dò Bất đẳng hướng Vi sóng Wilkinson) về mật độ baryon vũ trụ là $9,9 \times 10^{-30}$ g/cm³ và giả định rằng 90% khối lượng là khí liên thiên hà, khoảng 15% khối lượng thiên hà là sao (khoảng 80% khối lượng vật chất baryon) và ngôi sao trung bình nặng khoảng 0,7 khối lượng mặt trời (Read và Trentham 2005; Carroll và Ostlie 2007).

16. Armstrong và Sandberg (2013).

17. Ngay cả ở 100% vận tốc ánh sáng (không thể đạt được đối với các vật thể có khối lượng nghỉ khác 0), số lượng thiên hà có thể tiếp cận chỉ khoảng 6×10^9 . (So sánh với Gott và cộng sự [2005] và Heyl [2005].) Chúng tôi giả định rằng sự hiểu biết hiện tại về vật lý có liên quan là chính xác. Thật khó có thể tin tưởng cao vào bất kỳ giới hạn trên nào, vì ít nhất có thể hình dung rằng một nền văn minh siêu trí tuệ có thể mở rộng phạm vi của nó theo một cách nào đó mà chúng ta nghĩ là bất khả về mặt vật lý (ví dụ, bằng cách chế tạo cỗ máy thời gian, tạo ra các vũ trụ đang lạm phát mới, hoặc các phương pháp khác chưa được hình dung).

18. Số lượng các hành tinh có thể ở được trong mỗi hệ sao hiện không chắc chắn, vì vậy đây chỉ là ước tính thô. Traub (2012) dự đoán rằng một phần ba các hệ sao trong các lớp phổ F, G hoặc K có ít nhất một hành tinh đá là có thể ở được; xem thêm Clavin (2012). Các hệ sao F G K chiếm khoảng 22,7% số hệ sao trong vùng lân cận mặt trời, cho thấy rằng 7,6% số hệ sao có hành tinh phù hợp với sự sống. Ngoài ra, có thể có các hành tinh có thể ở được xung quanh

nhiều ngôi sao phổ M hơn (Gilster 2012). Xem thêm Robles và cộng sự (2008).

Không cần thiết phải đưa cơ thể con người vào sự khắc nghiệt của các chuyến du hành giữa các thiên hà. AI có thể giám sát quá trình thuộc địa hóa. Homo sapiens có thể được mang theo như thông tin, mà AI sau này có thể sử dụng để khởi tạo mẫu vật của loài chúng ta. Ví dụ, thông tin di truyền có thể được tổng hợp thành ADN và một thể hệ đầu tiên của con người có thể được thụ thai, nuôi dưỡng và giáo dục bởi những người bảo vệ AI có dáng vẻ như con người.

19. O'Neill (1974).

20. Dyson (1960) tuyên bố đã có được ý tưởng cơ bản từ nhà văn khoa học viễn tưởng Olaf Stapledon (1937), người có thể đã được truyền cảm hứng từ những suy nghĩ tương tự của J. D. Bernal (Dyson 1979, 211).

21. Nguyên lý của Landauer nói rằng có một lượng năng lượng tối thiểu cần thiết để thay đổi một bit thông tin, được gọi là giới hạn Landauer, bằng $kT \ln 2$, trong đó k là hằng số Boltzmann ($1,38 \times 10^{-23}$ J/K) và T là nhiệt độ. Nếu chúng ta giả sử mạch được duy trì ở mức khoảng 300 K, thì 1026 watt cho phép chúng ta xóa khoảng 1047 bit mỗi giây. (Về hiệu quả có thể đạt được của các thiết bị tính toán cơ học nano, xem Drexler [1992]. Xem thêm Bradbury [1999]; Sandberg [1999]; irković [2004]. Các nền tảng của nguyên lý Landauer vẫn còn gây tranh cãi; 2011].)

22. Các ngôi sao khác nhau về sản lượng năng lượng của chúng, nhưng Mặt trời là một ngôi sao dãy chính khá điển hình.

23. Một phân tích chi tiết hơn có thể xem xét kỹ hơn các loại tính toán mà chúng ta quan tâm. Số lượng tính toán tuần tự có thể được thực hiện khá hạn chế, vì một máy tính tuần tự nhanh phải nhỏ để giảm thiểu độ trễ giao tiếp trong các phần khác nhau của máy tính. Cũng có giới hạn về số lượng bit có thể được lưu trữ, và, như chúng ta đã thấy, về số bước không thể đảo ngược (liên quan đến việc xóa thông tin) có thể được thực hiện.

24. Điều đó có nghĩa là, khoảng cách sẽ là nhỏ đối với một số chỉ số của tự nhiên, chẳng hạn như logarit của quy mô dân số có thể được hỗ trợ bền vững ở mức sinh hoạt theo mức độ khả năng nhất định nếu tất cả các nguồn lực được dành cho mục đích đó.

25. Chúng tôi giả định ở đây rằng không có nền văn minh ngoài Trái đất nào có thể cản trở. Chúng tôi cũng giả định rằng giả thuyết mô phỏng là sai. Xem Bostrom (2003a). Nếu một trong những giả định này là không chính xác, có thể có những rủi ro phi nhân tạo quan trọng (những rủi ro liên quan đến tác tử thông minh không phải con người). Xem thêm Bostrom (2003b, 2009c).

26. Ít nhất một thể đơn nhất thông thái nắm bắt được ý tưởng về tiến hóa về nguyên tắc có thể tham gia một chương trình ưu sinh mà thông qua đó nó có thể từ từ nâng cao mức độ trí tuệ tập thể.

27. Tetlock và Belkin (1996).

28. Để rõ ràng: thuộc địa hóa và tái thiết kế một phần lớn của vũ trụ có thể tiếp cận hiện không nằm trong tầm với trực tiếp của loài người. Thuộc địa hóa liên thiên hà vượt xa công nghệ ngày nay. Vấn đề là về nguyên tắc, chúng ta có thể sử dụng các khả năng hiện tại

của mình để phát triển các khả năng bổ sung cần thiết, và do vậy, vấn đề này vẫn thuộc tầm với gián tiếp. Tất nhiên cũng đúng là nhân loại hiện không phải là một thể đơn nhất và chúng ta không biết rằng mình sẽ không bao giờ phải đối mặt với sự chống đối thông minh từ một thể lực bên ngoài nào đó nếu bắt đầu tái thiết kế phần vũ trụ có thể tiếp cận. Tuy nhiên, để đáp ứng ngưỡng bền vững của thể đơn nhất thông thái, sẽ là đủ khi ai đó sở hữu một tập hợp năng lực sao cho nếu một thể đơn nhất thông thái không đối mặt với sự chống đối thông minh đã sở hữu khả năng này thì sự xâm chiếm và tái cấu trúc của một phần lớn vũ trụ có thể tiếp cận sẽ nằm trong tầm với gián tiếp của nó.

29. Đôi khi có thể hữu ích khi nói về hai AI sở hữu những siêu năng lực đặc thù của riêng mình. Theo nghĩa mở rộng của từ này, người ta có thể hình dung một siêu cường như một thứ mà một tác tử sở hữu liên quan đến một lĩnh vực hành động nào đó. (Trong ví dụ này, có lẽ là một lĩnh vực bao gồm tất cả nền văn minh của loài người nhưng loại trừ AI kia.)

CHƯƠNG 7: Ý CHÍ CỦA SIÊU TRÍ TUỆ

1. Tất nhiên, điều này không thể phủ nhận rằng sự khác biệt có vẻ ngoài không đáng kể có thể có chức năng sâu sắc.

2. Yudkowsky (2008a, 310).

3. David Hume, nhà triết học Khai sáng Scotland, nghĩ rằng chỉ có niềm tin (chẳng hạn như về điều nên làm) thì không thể thúc đẩy hành động: mong muốn ở một mức độ nào đó là cần thiết. Điều này sẽ hỗ trợ cho luận đề trực giao với một sự phản đối khả thể, cụ thể là trí tuệ đầy đủ có thể đòi hỏi một số niềm tin nhất định mà sau đó nhất thiết sẽ tạo ra những động lực nhất định. Tuy nhiên, mặc dù luận đề

trực giao có thể thu hút sự ủng hộ từ lý thuyết động lực của Hume, nhưng nó không phải là tiền đề cho lý thuyết này. Cụ thể, người ta không cần bảo thủ rằng chỉ có niềm tin thì không bao giờ có thể tạo động lực cho hành động. Chẳng hạn, có thể giả định rằng một tác tử thông minh hết mức có thể được tạo động lực để theo đuổi bất kỳ hành động nào nếu tác tử đó có mong muốn nhất định về một sức mạnh vượt trội. Một cách khác trong đó luận đề trực giao có thể đúng ngay cả khi lý thuyết động lực của Hume là sai là: Trí tuệ cao tùy ý không đòi hỏi phải có được bất kỳ niềm tin nào (được cho là) về bản thân việc tạo động lực. Cách thứ ba mà theo đó có thể cho luận đề trực giao đúng ngay cả khi lý thuyết của Hume sai là: Xây dựng được một tác tử (hoặc trung lập hơn, một “quá trình tối ưu hóa”) với trí tuệ cao tùy ý nhưng có thể chế xa lạ đến mức không có bất kỳ chức năng nào tương tự như những gì mà con người gọi là “niềm tin” và “mong muốn. (Xem một số nỗ lực gần đây để bảo vệ lý thuyết động lực của Hume trong Smith [1987], Lewis [1988] và Sinhababu [2009].)

4. Chẳng hạn, Derek Parfit đã lập luận rằng một số ưu tiên cơ bản nhất định sẽ là không hợp lý, chẳng hạn như của một tác tử bình thường khác với một “Sự cào bằng các ngày thứ ba trong tương lai”:

Một người theo chủ nghĩa khoái lạc rất quan tâm về chất lượng của những trải nghiệm trong tương lai của mình. Với một ngoại lệ, anh ta có mối quan tâm như nhau với mọi thành tố cấu thành nên tương lai của mình. Anh ta có một “Sự cào bằng các ngày thứ ba trong tương lai”. Với bất kỳ ngày thứ ba nào, anh ta đều quan tâm theo cách thông thường về những gì đang xảy ra với mình. Nhưng anh ta không bao giờ quan tâm đến những nỗi đau hay thú vui có thể xảy ra về một Thứ ba trong tương lai... Sự thờ ơ này hoàn toàn là

thật. Khi lên kế hoạch cho tương lai của mình, đúng là anh ta luôn thích viễn cảnh đau khổ vào thứ ba hơn những sự đau khổ nhẹ nhất vào bất kỳ ngày nào khác. [Parfit (1986, 123-124); xem thêm Parfit (2011)]

Với mục đích của mình, chúng ta không cần biết liệu Parfit có đúng về việc tác tử này là phi lý do (irrational) hay không, miễn là chúng ta cho rằng anh ta không nhất thiết là không thông minh theo nghĩa công cụ được giải thích trong sách. Tác tử của Parfit có thể có lý do công cụ hoàn hảo, và do đó có một trí tuệ tuyệt vời, ngay cả khi anh ta thiếu đi một sự nhạy cảm nào đó với “lý do khách quan”, thứ cần có để trở thành một tác tử có lý do hoàn chỉnh. Do đó, loại ví dụ này không phá hỏng luận đề trực giao.

5. Ngay cả khi có những sự thật đạo đức khách quan mà bất kỳ tác tử có lý do nào cũng có thể hiểu được, và ngay cả khi những sự thật đạo đức này bằng cách nào đó về bản chất có thể tạo ra động lực (để bất kỳ ai hiểu được chúng đều nhất thiết phải có động lực để hành động theo chúng), thì điều này không hẳn phải phá hỏng luận đề trực giao. Luận đề này vẫn có thể đúng nếu một tác tử có thể có lý do công cụ hoàn hảo ngay cả khi thiếu một số năng lực khác cấu thành lý do, hoặc một số năng lực cần để hiểu đầy đủ các sự kiện đạo đức khách quan. (Một tác tử cũng có thể cực kỳ thông minh, thậm chí có siêu trí tuệ, mà không có lý do công cụ đầy đủ trong mọi lĩnh vực.)

6. Xem thêm về luận đề trực giao trong Bostrom (2012) và Armstrong (2013).

7. Sandberg và Bostrom (2008).

8. Stephen Omohundro đã viết hai bài báo tiên phong về chủ đề này (Omohundro 2007, 2008). Omohundro lập luận rằng tất cả các hệ thống AI tiên tiến có khả năng thể hiện một số “động lực cơ bản” (basic drive), có nghĩa là “xu hướng sẽ xuất hiện trừ khi bị chống đối rõ ràng”. Thuật ngữ “động lực AI” có một ưu điểm là ngắn gọn và gợi mở, nhưng nhược điểm nằm ở chỗ nó cho thấy các mục tiêu công cụ mà nó đề cập ảnh hưởng đến việc ra quyết định của AI giống như các động lực tâm lý ảnh hưởng đến việc ra quyết định của con người, nghĩa là thông qua một loại “lực kéo” theo hiện tượng luận (tức là có liên quan đến trải nghiệm và ý thức) lên cái tôi của chúng ta, và sức mạnh ý chí của chúng ta đôi khi có thể thành công trong việc chống lại nó. Khái niệm này không giúp ích gì nhiều. Mọi người thường không nói rằng một người bình thường có một “động lực” để điền vào tờ khai thuế của họ, mặc dù việc nộp thuế có thể là một mục tiêu tương đối hội tụ cho con người trong xã hội đương đại (một mục tiêu mà việc thực hiện sẽ tránh được rắc rối vốn sẽ ngăn chúng ta thực hiện nhiều mục tiêu tối thượng). Cách xử lý của chúng ta ở đây cũng khác với Omohundro ở một số điểm khác quan trọng hơn, mặc dù ý tưởng cơ bản là như nhau. (Xem thêm Chalmer [2010] và Omohundro [2012].)

9. Chislenko (1997).

10. Xem thêm Shulman (2010b).

11. Một tác tử cũng có thể thay đổi phép biểu diễn mục tiêu của nó, nếu bản thể học của nó thay đổi, để chuyển phép biểu diễn cũ của nó sang bản thể học mới; xem de Blanc (2011).

Một loại yếu tố khác có thể khiến một tác tử theo lý thuyết quyết định bằng chứng thực hiện nhiều hành động khác nhau, bao gồm cả

việc thay đổi mục tiêu tối thượng, là nhập bằng chứng cho việc quyết định làm như vậy. Ví dụ, một tác tử theo lý thuyết quyết định bằng chứng có thể tin rằng có tồn tại các tác tử khác giống như nó trong vũ trụ, và hành động của chính nó sẽ cung cấp một số bằng chứng về cách các tác tử khác hành động. Do đó, tác tử này có thể chọn áp dụng mục tiêu tối thượng có tính vị tha đối với các tác tử liên kết theo bằng chứng khác, với lý do điều này sẽ cung cấp bằng chứng cho nó rằng các tác tử khác sẽ chọn hành động như vậy. Tuy nhiên, một kết quả tương đương có thể đạt được mà không cần thay đổi mục tiêu tối thượng, bằng cách chọn từng mục tiêu để hành động như là một mục tiêu tối thượng.

12. Một tài liệu tâm lý sâu rộng khám phá sự hình thành sở thích thích nghi.

Ví dụ, xem Forgas và cộng sự (2010).

13. Trong các mô hình chính thức, giá trị của thông tin được định lượng là sự khác biệt giữa giá trị mong đợi được hiện thực hóa bởi các quyết định tối ưu được thực hiện với thông tin đó và giá trị mong đợi được hiện thực hóa bằng các quyết định tối ưu được thực hiện mà không có nó. (Ví dụ, xem Russell và Norvig [2010].) Theo đó, giá trị của thông tin không bao giờ là âm. Nó cũng kéo theo việc bất kỳ thông tin nào mà bạn biết rằng sẽ không bao giờ ảnh hưởng đến các quyết định của bạn sẽ có giá trị bằng không đối với bạn. Tuy nhiên, loại mô hình này giả định một vài sự lý tưởng hóa thường không phù hợp trong thế giới thực, chẳng hạn như kiến thức đó không có giá trị cuối cùng (có nghĩa là kiến thức chỉ có giá trị công cụ và không có giá trị vì lợi ích của chính nó) và các tác tử không minh bạch với các tác tử khác.

14. Ví dụ, Hájek (2009).

15. Chiến lược này được minh họa bằng việc ấu trùng mực biển bơi xung quanh cho đến khi tìm thấy một tảng đá phù hợp sau đó sẽ tự gắn vào nó vĩnh viễn. Được gắn kết vào đúng chỗ, ấu trùng ít có nhu cầu xử lý thông tin phức tạp, do đó nó tiến hành tiêu hóa một phần não của chính mình (hạch não). Người ta có thể quan sát hiện tượng tương tự trong một số học giả khi họ đã có được chức vụ.

16. Bostrom (2012).

17. Bostrom (2006c).

18. Thay vào đó, người ta có thể đảo ngược câu hỏi và xem xét các lý do có thể khiến một thể đơn nhất siêu trí tuệ không phát triển một số khả năng công nghệ. Các lý do này gồm: (a) thể đơn nhất thấy trước rằng khả năng đó không có tác dụng; (b) chi phí phát triển quá lớn so với độ thỏa dụng dự đoán của nó (ví dụ: nếu công nghệ sẽ không bao giờ phù hợp để đạt được bất kỳ kết thúc đơn lẻ nào, hoặc nếu thể đơn nhất có tỷ lệ chiết khấu rất cao làm nản lòng các nhà đầu tư); (c) thể đơn nhất có một số giá trị cuối cùng đòi hỏi phải tránh các con đường phát triển công nghệ cụ thể; (d) nếu thể đơn nhất không chắc chắn rằng bản thân sẽ vẫn ổn định, nó sẽ không muốn phát triển các công nghệ có thể đe dọa sự ổn định nội bộ của mình hoặc làm cho hậu quả của việc giải thể trở nên tồi tệ hơn (ví dụ, chính phủ thế giới có thể không muốn phát triển công nghệ tạo điều kiện cho sự nổi loạn, ngay cả khi chúng có một số ứng dụng tốt, cũng như không phát triển các công nghệ để dễ dàng sản xuất vũ khí hủy diệt hàng loạt có thể tàn phá nếu chính phủ thế giới giải thể); (e) tương tự, thể đơn nhất có thể đã thực hiện một số loại cam kết chiến lược ràng buộc không phát triển một số công nghệ nhất định, có tác

dụng ngay cả khi điều kiện hiện tại có thể thuận tiện để phát triển công nghệ đó. (Tuy nhiên, lưu ý rằng một số lý do hiện tại để phát triển công nghệ sẽ không áp dụng cho thể đơn nhất: ví dụ: lý do phát sinh từ các cuộc đua vũ trang.)

19. Giả sử một tác tử giảm lượng tài nguyên thu được trong tương lai với tốc độ theo cấp số nhân, và do giới hạn tốc độ ánh sáng, nó chỉ có thể tăng tài nguyên của mình với tốc độ như một hàm đa thức. Liệu điều này có nghĩa là sẽ có một thời gian sau đó các tác tử sẽ không thấy đáng để tiếp tục mở rộng quá trình thụ đắc tài nguyên? Không, vì mặc dù giá trị hiện tại của các tài nguyên thu được ở thời điểm trong tương lai sẽ càng tiệm cận với 0 khi chúng ta nhìn càng xa vào tương lai, và chi phí hiện tại để có được chúng cũng vậy. Chi phí hiện tại của việc gửi thêm một “tàu von Neumann” 100 triệu năm kể từ bây giờ (có thể sử dụng một số tài nguyên có được một thời gian ngắn trước đó) sẽ bị giảm theo cùng một hệ số giảm mà theo đó, giá trị hiện tại của các tài nguyên trong tương lai mà tàu đó thu được sẽ giảm (nếu hệ số đó không đổi).

20. Mặc dù thể tích không gian đạt được bằng các tàu thuộc địa hóa tại một thời điểm nhất định có thể gần như hình cầu và mở rộng ở tốc độ tỷ lệ thuận với bình phương thời gian trôi qua kể từ khi phóng con tàu đầu tiên ($\sim t^2$), lượng tài nguyên chứa trong thể tích không gian này sẽ tuân theo một mô hình tăng trưởng kém đều đặn hơn, vì sự phân phối tài nguyên là không đồng nhất và khác nhau giữa một số quy mô. Ban đầu, tốc độ tăng trưởng có thể là $\sim t^2$ khi hành tinh chủ bị thuộc địa hóa; sau đó tốc độ tăng trưởng có thể trở nên đột biến khi các hành tinh và hệ mặt trời gần đó bị thuộc địa hóa; sau đó, khi thể tích không gian gần như hình đĩa của Dải Ngân hà

được lấp đầy, tốc độ tăng trưởng sẽ chững lại, xấp xỉ tỷ lệ thuận với t ; sau đó nó có thể lại trở nên đột biến khi các thiên hà gần đó bị thuộc địa hóa; kế tiếp, nó có thể một lần nữa xấp xỉ $\sim t^2$ khi quá trình mở rộng diễn ra theo quy mô mà sự phân bố của các thiên hà gần như đồng nhất; sau đó là một giai đoạn tăng trưởng đột biến khác theo sau là sự tăng trưởng $\sim t^2$ ổn định khi các siêu cụm thiên hà bị thuộc địa hóa; cho đến khi rốt cuộc tốc độ tăng trưởng bắt đầu một sự suy giảm cuối cùng về 0 khi tốc độ giãn nở của vũ trụ tăng đến mức khiến cho việc thuộc địa hóa xa hơn là không thể.

21. Quan điểm mô phỏng có thể đặc biệt quan trọng trong bối cảnh này. Một tác tử siêu trí tuệ có thể chỉ định một xác suất đáng kể cho các giả thuyết mà theo đó, nó sống trong một bản mô phỏng máy tính và trình tự đối tượng tri giác của nó được tạo ra bởi một siêu trí tuệ khác, và điều này có thể tạo ra các lý do công cụ hội tụ khác nhau tùy thuộc vào phỏng đoán của tác tử về loại mô phỏng nào nó có nhiều khả năng “cư ngụ” nhất. So sánh với Bostrom (2003a).

22. Khám phá các định luật cơ bản của vật lý và các sự kiện cơ bản khác về thế giới là một mục tiêu công cụ hội tụ. Ở đây, chúng ta có thể đặt nó dưới dòng chữ “tăng cường nhận thức” được viết hoa in đậm, tuy nó cũng có thể phát sinh từ mục tiêu “hoàn thiện công nghệ” (vì các hiện tượng vật lý mới có thể tạo điều kiện ra đời các công nghệ mới).

CHƯƠNG 8: KẾT QUẢ MẶC ĐỊNH CÓ PHẢI LÀ NGÀY TẬN THẾ?

1. Chúng ta có thêm một số rủi ro diệt vong khác ở các kịch bản mà trong đó, nhân loại sống sót trong một trạng thái kém xa mức tối ưu, hoặc trong đó, một phần lớn tiềm năng phát triển mong muốn của

chúng ta bị lãng phí không thể đảo ngược. Trên hết, có thể có những rủi ro diệt vong liên quan đến việc dẫn đến một sự bùng nổ trí tuệ tiềm năng, ví dụ, phát sinh từ chiến tranh giữa các quốc gia cạnh tranh phát triển siêu trí tuệ.

2. Có một thời điểm dễ tổn hại quan trọng khi AI lần đầu tiên nhận ra sự cần thiết của việc che giấu như vậy (một sự kiện mà chúng ta có thể gọi là sự hình thành của ý định lừa dối). Bản thân nhận thức ban đầu này sẽ không được cố tình che giấu khi nó xuất hiện. Nhưng khi đã có nhận thức này, AI có thể biến chuyển nhanh chóng để che giấu việc nó đã xuất hiện, trong khi thiết lập một số động lực nội bộ bí mật (có thể được ngụy trang thành một quá trình vô hại hòa trộn với tất cả các quá trình phức tạp khác đang diễn ra trong tâm trí của nó) sẽ cho phép nó tiếp tục hoạch định chiến lược dài hạn một cách kín đáo.

3. Ngay cả các tin tặc cũng có thể viết những chương trình nhỏ và dường như vô hại để thực hiện những điều hoàn toàn bất ngờ. (Ví dụ: xem một số bài thi giành chiến thắng trong International Obfuscated C Code Contest.)

4. Điểm mà tại đó một số biện pháp kiểm soát AI có thể hiệu quả trong một bối cảnh cố định nhưng lại thất bại thảm hại khi bối cảnh thay đổi cũng được Eliezer Yudkowsky nhấn mạnh; Ví dụ, xem Yudkowsky (2008a)

5. Thuật ngữ này dường như được đặt ra bởi nhà văn khoa học viễn tưởng Larry Niven (1973), nhưng dựa trên các thí nghiệm thường kích thích não trong thế giới thực; so sánh với Olds và Milner (1954) và Oshima và Katayama (2010). Xem thêm Ring và Orseau (2011).

6. Bostrom (1997).

7. Có thể có một số dạng thực thi khả thể của một cơ chế học tăng cường dẫn đến việc mất khả năng an toàn thay vì dẫn đến dư thừa hạ tầng khi AI phát hiện ra giải pháp wireheading. Vấn đề là việc này có thể dễ dàng đi sai và thất bại vì lý do bất ngờ.

8. Điều này đã được Marvin Minsky đề xuất (tham khảo Russell và Norvig [2010, 1039]).

9. Vấn đề về loại tâm trí kỹ thuật số nào sẽ có ý thức, theo nghĩa là có trải nghiệm chủ quan (hay “qualia” theo ngôn ngữ triết học) là rất quan trọng khi nhắc đến đến điểm này (mặc dù nó không liên quan đến nhiều phần khác trong sách). Một câu hỏi mở là việc ước tính chính xác cách một thực thể giống con người hành xử trong nhiều hoàn cảnh khác nhau mà không mô phỏng bộ não của nó đủ chi tiết để mang ý thức sẽ khó khăn ra sao. Một câu hỏi khác là liệu có các thuật toán sử dụng chung cho siêu trí tuệ (ví dụ như các kỹ thuật học tăng cường) sao cho việc thực thi chúng sẽ tạo ra qualia hay không. Ngay cả khi chúng ta đánh giá xác suất của việc bất kỳ chương trình con nào như vậy sẽ có ý thức khá nhỏ thì số bản sao của chúng có thể lớn đến mức thậm chí một rủi ro nhỏ mà chúng có thể gặp phải cũng sẽ có trọng số đáng kể trong quá trình tính toán đạo đức của chúng ta. Xem thêm Metzinger (2003, Chương 8).

10. Bostrom (2002a, 2003a); Elga (2004).

CHƯƠNG 9: VẤN ĐỀ KIỂM SOÁT

1. Ví dụ: Laffont và Martimort (2002).

2. Giả sử đa số cử tri muốn đất nước của họ xây dựng một số loại siêu trí tuệ đặc biệt. Họ bầu một ứng cử viên hứa hẹn sẽ thực hiện, nhưng có thể gặp khó khăn để đảm bảo rằng ứng cử viên đó,

một khi nắm quyền, sẽ làm như đã hứa trong chiến dịch tranh cử và theo đuổi dự án theo cách mà cử tri mong đợi. Giả sử người này làm đúng với lời nói của mình và chỉ đạo chính phủ ký hợp đồng với một tổ chức học thuật hoặc công nghiệp để thực hiện công việc; nhưng một lần nữa lại có những vấn đề đại diện: các quan chức trong chính phủ có thể có quan điểm riêng của họ về những gì nên làm, và có thể thực hiện dự án theo cách tôn trọng câu chữ nhưng không phải là tinh thần của chỉ thị từ người lãnh đạo. Ngay cả khi chính phủ thực hiện công việc của mình một cách trung thực, các đối tác khoa học có hợp đồng có thể cũng có kế hoạch riêng. Vấn đề này tái diễn trên nhiều cấp độ. Giám đốc của một trong những phòng thí nghiệm tham gia có thể tỉnh táo lo lắng về việc một kỹ thuật viên đưa vào bản thiết kế một yếu tố không được phê chuẩn (hãy tưởng tượng Tiến sĩ T.R. Eason lén vào văn phòng của mình vào một đêm khuya, đăng nhập vào cơ sở mã dự án và viết lại một phần hệ thống mục tiêu của AI hạt giống). AI này sẽ nói “phục vụ Tiến sĩ Eason”, trong khi đáng ra phải nói “phục vụ nhân loại”.

3. Tuy nhiên, ngay cả đối với việc phát triển siêu trí tuệ, có thể có một người giữ vai trò thử nghiệm hành vi với tư cách là một yếu tố phụ trợ trong một loạt các biện pháp an toàn. Mặc dù vậy, nếu một AI hoạt động sai trong giai đoạn phát triển (mà đây là một việc rất tệ), điều ngược lại sẽ không đúng.

4. Trong một ví dụ kinh điển từ năm 1975, Steven Dompier đã viết một chương trình cho Altair 8800 tận dụng hiệu ứng này (và sự thiếu che chắn xung quanh phần vỏ của chiếc máy tính siêu nhỏ). Việc chạy chương trình sẽ phát ra các sóng điện từ tạo ra âm nhạc khi người ta cầm một đài phát thanh bán dẫn đứng gần máy tính

(Driscoll 2012). Bill Gates khi còn trẻ đã tham dự một buổi trình diễn và nói rằng ông đã bị trò này làm cho ấn tượng và hoang mang (Gates 1975). Trong mọi trường hợp đều có kế hoạch thiết kế chip trong tương lai với khả năng Wi-Fi tích hợp (Greene 2012).

5. Việc có một niềm tin mà nếu chúng ta có cơ hội hành động theo nó, có thể dẫn đến sự hủy hoại đối với toàn bộ tài nguyên vũ trụ không phải là vấn đề nhỏ. Người ta có thể tranh luận về nguyên tắc sau: nếu trước đây ai đó có thể khẳng định về N lần mà trong đó một hệ thống đã được cải thiện đủ để đảm bảo an toàn, và mỗi lần nó được tiết lộ rằng họ đã sai, thì trong lần tiếp theo họ không có quyền gán tín nhiệm lớn hơn $1/(N + 1)$ cho hệ thống được an toàn.

6. Trong một thí nghiệm không chính thức, một người thông minh đã đóng vai một AI. Một cá nhân khác đóng vai trò là người gác cổng và được giao nhiệm vụ không cho AI ra khỏi hộp. AI có thể liên lạc với người gác cổng chỉ bằng văn bản và được dành hai giờ để thuyết phục người gác cổng cho phép. Trong ba trên năm trường hợp, với các cá nhân khác nhau đóng vai người gác cổng, AI đã trốn thoát (Yudkowsky 2002). Những gì một con người có thể làm, một siêu trí tuệ cũng có thể. (Tất nhiên, điều ngược lại là không đúng. Ngay cả khi nhiệm vụ cho một siêu trí tuệ thực sự khó khăn hơn, và có lẽ những người gác cổng sẽ có động lực mạnh mẽ hơn để kiểm chế việc phát hành AI hơn những người chơi gác cổng trong thí nghiệm, siêu trí tuệ vẫn có thể thành công trong khi con người sẽ thất bại.)

7. Người ta không nên nói quá về mức an toàn cận biên có thể đạt được theo cách này. Hình ảnh tinh thần có thể thay thế cho màn

hình đồ họa. Hãy xem xét những tác động mà sách có thể có đối với mọi người, và những cuốn sách thậm chí còn không tương tác.

8. Xem thêm Chalmer (2010). Sẽ là một sai lầm khi suy luận rằng việc xây dựng một hệ thống không bao giờ được quan sát bởi bất kỳ thực thể bên ngoài nào là không thể có tác dụng. Người ta có thể đặt một giá trị cuối cùng vào những gì diễn ra trong một hệ thống như vậy. Ngoài ra, những người khác có thể có sở thích về những gì diễn ra trong một hệ thống như vậy, và do đó có thể bị ảnh hưởng bởi sự sáng tạo của nó hoặc lời hứa về sự sáng tạo đó. Kiến thức về sự tồn tại của một số loại hệ thống bị cô lập (những hệ thống có chứa người quan sát) cũng có thể gây ra sự không chắc chắn mang tính vị nhân ở người quan sát bên ngoài. Điều này có thể ảnh hưởng đến hành vi của họ.

9. Người ta có thể tự hỏi tại sao hội nhập xã hội được coi là một hình thức kiểm soát năng lực. Thay vào đó, nó không nên được phân loại như một phương pháp lựa chọn động lực trên cơ sở mà nó liên quan đến việc tìm cách ảnh hưởng đến một hành vi của hệ thống bằng các biện pháp khuyến khích? Chúng ta sẽ xem xét kỹ việc lựa chọn động; nhưng, để trả lời cho câu hỏi này, chúng ta đang coi lựa chọn động lực như một cụm phương pháp kiểm soát hoạt động bằng cách chọn hoặc định hình một hệ thống mục tiêu tối thượng (các mục tiêu nhằm tới mục đích của chính chúng thay vì chỉ là lý do công cụ). Hội nhập xã hội không nhằm vào một mục tiêu tối thượng của hệ thống, vì vậy nó không phải là lựa chọn động lực. Thay vào đó, hội nhập xã hội nhằm tới việc hạn chế các khả năng hiệu quả của hệ thống: nó tìm cách khiến hệ thống không có khả năng đạt được một tập hợp kết quả nhất định, những kết quả mà trong đó, hệ thống đạt

được lợi ích của việc “đào tẩu” mà không phải chịu các hình phạt liên quan (bị trừng phạt và mất đi lợi ích của việc hợp tác). Hy vọng rằng bằng cách hạn chế kết quả mà hệ thống có thể đạt được, hệ thống sẽ thấy rằng phương tiện hiệu quả nhất còn lại để thực hiện các mục tiêu tối thượng của nó là hành xử hợp tác.

10. Cách tiếp cận này có thể sẽ hứa hẹn hơn trong trường hợp một bản giả lập được cho là có động lực giống con người.

11. Tôi nợ Carl Shulman ý tưởng này.

12. Tạo một mật mã nhất định qua mặt được trình phá mã siêu trí tuệ là một thách thức không nhỏ. Ví dụ, dấu vết của các số ngẫu nhiên có thể được để lại trong não của một người quan sát hoặc trong cấu trúc vi mô của trình tạo mã ngẫu nhiên mà từ đó, siêu trí tuệ có thể trích xuất; hoặc, nếu các số giả ngẫu nhiên được sử dụng, siêu trí tuệ có thể đoán hoặc khám phá ra phần cốt lõi mà từ đó chúng được tạo ra. Hơn nữa, siêu trí tuệ có thể xây dựng các máy tính lượng tử lớn, hoặc thậm chí khám phá các hiện tượng vật lý chưa biết đến, có thể được sử dụng để xây dựng các loại máy tính mới.

13. AI có thể tự đánh tín hiệu cho mình để tin rằng mình đã nhận được token phần thưởng, nhưng điều này không nên biến nó thành wirehead nếu nó được thiết kế để muốn token phần thưởng (trái ngược với việc muốn ở trong trạng thái có niềm tin nhất định về token phần thưởng).

14. Xem bài viết gốc trong Bostrom (2003a). Xem thêm Elga (2004).

15. Shulman (2010a).

16. Thực tại nền tảng có lẽ chứa nhiều tài nguyên tính toán hơn thực tế mô phỏng, vì bất kỳ quá trình tính toán nào xảy ra trong một mô phỏng cũng xảy ra trên máy tính chạy mô phỏng. Thực tại nền tảng cũng có thể chứa vô số tài nguyên vật lý khác, điều mà các tác tử mô phỏng khó có thể truy cập – các tác tử chỉ tồn tại ở sự nuông chiều của các trình giả lập mạnh mẽ, vốn có thể có mục đích sử dụng khác đối với các tài nguyên đó. (Tất nhiên, suy luận ở đây không hoàn toàn có giá trị về mặt suy diễn: về nguyên tắc, có thể các vũ trụ mà trong đó, bản mô phỏng được chạy chứa nhiều tài nguyên mà các nền văn minh mô phỏng về trung bình có quyền truy cập vào nhiều tài nguyên hơn các nền văn minh phi mô phỏng, thậm chí mặc dù mỗi nền văn minh phi mô phỏng chạy mô phỏng có nhiều tài nguyên hơn tất cả các nền văn minh mà nó mô phỏng kết hợp lại.)

17. Có nhiều ý kiến cân nhắc khác có thể có liên quan đến vấn đề này, mà những hệ quả của chúng chưa được phân tích đầy đủ. Những ý kiến cân nhắc này rất cuộc có thể rất quan trọng trong việc phát triển một phương pháp được coi là tất cả mọi thứ để đối phó với viễn cảnh xảy ra một sự bùng nổ trí tuệ. Tuy nhiên, dường như chúng ta sẽ không thành công trong việc tìm ra nguồn thực của những lập luận đó trừ khi chúng ta lần đầu tiên đạt được một số tiến bộ trong các ý kiến cân nhắc “trần tục” hơn, vốn là chủ đề của hầu hết nội dung trong cuốn sách này.

18. Ví dụ, so sánh với Quine và Ullian (1978).

19. Thứ mà AI có thể điều tra bằng cách xem xét các đặc tính hiệu suất của các chức năng tính toán cơ bản khác nhau, chẳng hạn như kích thước và dung lượng của các bus dữ liệu khác nhau, thời

gian cần thiết để truy cập các phần khác nhau của bộ nhớ, tỷ lệ lật bit ngẫu nhiên và những giá trị tương tự.

20. Tiên nghiệm có thể là (một giá trị xấp xỉ tính toán được) của tiên nghiệm Solomonoff (tiên nghiệm này gán xác suất cho các thế giới khả thể trên cơ sở độ phức tạp thuật toán của chúng). Xem Li và Vitányi (2008).

21. Ở khoảnh khắc sau sự nhận thức về sự lừa dối, AI có thể sẽ cố gắng xóa dấu vết của ý nghĩ đột biến của nó. Do đó, điều quan trọng là dây bẫy (tripwire) này hoạt động liên tục. Nó cũng sẽ là một thực hành tốt khi sử dụng một máy ghi âm trên máy bay, lưu trữ dấu vết hoàn chỉnh của tất cả hoạt động của AI (bao gồm cả thời gian chính xác của tín hiệu đầu vào qua bàn phím từ các lập trình viên), để quỹ đạo của nó có thể được lưu lại hoặc phân tích sau khi tự động tắt. Thông tin có thể được lưu trữ trên phương tiện ghi một lần đọc nhiều lần.

22. Asimov (1942). Sau đó, một “Điều luật 0” đã được bổ sung: (0) Một robot có thể không gây hại cho nhân loại, hoặc, bằng cách không hành động, bỏ mặc loài người gặp phải hiểm nguy (Asimov 1985).

23. So sánh với Gunn (1982).

24. Russell (1986, 161f).

25. Tương tự, mặc dù một số nhà triết học đã dành toàn bộ sự nghiệp để cố gắng xây dựng một cách cẩn thận các hệ thống phi thần học, các trường hợp và hậu quả mới đôi khi được phát hiện là cần khắc phục. Ví dụ, sự phân biệt giữa các hậu quả dự định và không lường trước được, và các vấn đề khác tương tự; ví dụ, xem Kamm (2007).

26. Armstrong (2010).

27. Theo nguyên tắc thông thường, nếu một người có kế hoạch sử dụng nhiều cơ chế an toàn để kiểm chế AI, thì có thể là khôn ngoan khi làm việc trên từng cơ chế như thể nó được dự định là cơ chế an toàn duy nhất và như thế vì thế mà nó được yêu cầu phải đầy đủ. Nếu người ta đặt một cái thùng bị rò rỉ bên trong một cái thùng bị rò rỉ khác, nước vẫn chảy ra.

28. Một biến thể của ý tưởng này là xây dựng AI để nó liên tục được tạo động lực hành động theo những dự đoán tốt nhất của nó về tiêu chuẩn được xác định ngầm. Trong cách thiết lập này, mục tiêu tối thượng của AI luôn là hành động theo tiêu chuẩn được xác định ngầm, và nó theo đuổi một cuộc điều tra về tiêu chuẩn này chỉ vì lý do công cụ.

CHƯƠNG 10: TIỀN TRI, THẦN ĐÈN, TOÀN NĂNG, CÔNG CỤ

1. Những cái tên này, tất nhiên, là giống với con người và ta không nên chỉ coi chúng là những thứ tương tự. Chúng chỉ có ý nghĩa như là nhãn cho một số khái niệm khác nhau được tạm coi là đúng về các loại hệ thống khả thể mà chúng ta có thể xem xét cố gắng xây dựng.

2. Về kết quả của cuộc bầu cử tiếp theo, người ta sẽ không muốn được trả lời bằng một danh sách tổng hợp về vị trí dự kiến và vector động lượng của các hạt gần đó.

3. Được chỉ biệt vào một tập lệnh cụ thể trên một máy cụ thể.

4. Kuhn (1962); de Blanc (2011).

5. Sẽ khó hơn khi áp dụng một “phương pháp đồng thuận” như vậy với AI thần đèn hoặc AI toàn năng, vì thường có thể có nhiều chuỗi hành động cơ bản (chẳng hạn như gửi các mẫu tín hiệu điện cho các bộ truyền động của hệ thống) sẽ gần như có hiệu quả hoàn

toàn tương đương trong việc đạt được một mục tiêu nhất định; Các tác tử hơi khác nhau có thể lựa chọn một cách hợp lệ các hành động hơi khác nhau, dẫn đến việc không đạt được sự đồng thuận. Ngược lại, với các câu hỏi được đặt ra một cách thích hợp, thường sẽ có một số lượng nhỏ các câu trả lời phù hợp (chẳng hạn như “có” và “không”). (Về khái niệm điểm Schelling, còn được gọi là “tiêu điểm”, xem Schelling [1980].)

6. Phải chăng nền kinh tế thế giới ở một số khía cạnh cũng tương tự như một vị “thần đèn” yếu đuối, tuy mệnh lệnh đưa ra sẽ bị tính phí? Một nền kinh tế lớn hơn nhiều, như có thể phát triển trong tương lai, sau đó có thể có sức mạnh tương đương một vị thần với trí tuệ siêu tập thể.

Một khía cạnh quan trọng, trong đó nền kinh tế hiện tại không giống như một vị thần là mặc dù có thể (có tính phí) ra lệnh cho nền kinh tế giao bánh pizza đến tận nhà mình, tôi lại không thể bảo nó mang lại hòa bình thế giới. Lý do không phải là nền kinh tế không đủ mạnh mẽ, mà là nó không được phối hợp đầy đủ. Về mặt này, nền kinh tế giống như một tập hợp các vị “thần đèn” phục vụ những người cầm đèn khác nhau (với các kế hoạch cạnh tranh) hơn là chỉ một vị thần hoặc bất kỳ loại tác tử hợp nhất nào khác. Tăng tổng sức mạnh của nền kinh tế bằng cách làm cho mỗi vị thần cấu thành mạnh hơn, hoặc bằng cách thêm nhiều vị thần khác, sẽ không nhất thiết làm cho nền kinh tế có khả năng mang lại hòa bình. Để hoạt động như một vị thần siêu trí tuệ, nền kinh tế không chỉ cần phát triển khả năng sản xuất hàng hóa và dịch vụ với giá cả phải chăng (bao gồm cả những thứ đòi hỏi công nghệ hoàn toàn mới), mà còn cần phải giải quyết vấn đề phối hợp toàn cầu tốt hơn.

7. Nếu vị thần này bằng cách nào đó không có khả năng chống lại mệnh lệnh tiếp theo, và bằng cách nào đó không có khả năng tự lập trình lại để thoát khỏi tính dễ bị thuyết phục này, nó có thể hành động để ngăn chặn việc bất kỳ mệnh lệnh mới nào được ban hành.

8. Ngay cả một AI tiên tri bị giới hạn trong việc đưa ra câu trả lời có/không cũng có thể được sử dụng để tạo điều kiện cho việc tìm kiếm một AI thần đèn hoặc AI toàn năng, hoặc thực sự có thể được sử dụng trực tiếp như một thành phần trong các AI như vậy. AI tiên tri cũng có thể được sử dụng để tạo mã thực tế cho một AI như vậy nếu có thể hỏi số lượng câu hỏi đủ lớn. Một loạt các câu hỏi như vậy có thể có dạng gần như sau: “Trong phiên bản nhị phân của mã AI đầu tiên mà bạn nghĩ sẽ tạo thành một AI thần đèn, biểu tượng thứ n có phải là số 0 hay không?”

9. Người ta có thể tưởng tượng một AI tiên tri hoặc thần đèn phức tạp hơn một chút chỉ chấp nhận các câu hỏi hoặc mệnh lệnh nếu chúng được ban hành bởi một cơ quan được chỉ định, mặc dù điều này vẫn để ngỏ khả năng cơ quan đó biến chất hoặc bị bên thứ ba tống tiền.

10. John Rawls, một nhà triết học chính trị hàng đầu của thế kỷ XX, đã sử dụng phương tiện mô tả của một bức màn vô thức như một cách mô tả các loại ưu tiên cần được tính đến khi xây dựng một kế ước xã hội. Rawls gợi ý chúng ta nên tưởng tượng rằng mình đang chọn một kế ước xã hội đằng sau bức màn vô thức đã ngăn cản chúng ta biết mình sẽ là ai và sẽ giữ vai trò nào trong xã hội, nghĩa là trong tình huống như vậy, chúng ta sẽ phải suy nghĩ xã hội nào nói chung sẽ công bằng nhất và có thể mong muốn nhất mà không liên quan đến lợi ích cá nhân và những thiên kiến vị kỷ có thể

khiến chúng ta ưu tiên một trật tự xã hội mà trong đó, bản thân chúng ta được hưởng những đặc quyền. Xem Rawls (1971).

11. Karnofsky (2012).

12. Một ngoại lệ có thể xảy ra là phần mềm được nối với các bộ truyền động đủ mạnh, chẳng hạn như phần mềm trong các hệ thống cảnh báo sớm nếu được kết nối trực tiếp với đầu đạn hạt nhân hoặc với các sĩ quan được ủy quyền để tiến hành một cuộc tấn công hạt nhân. Trục trặc trong các phần mềm như vậy có thể dẫn đến các tình huống rủi ro cao. Điều này đã xảy ra ít nhất hai lần. Ngày 9 tháng 11 năm 1979, một sự cố máy tính đã khiến NORAD (Bộ Tư lệnh Phòng thủ Không gian Bắc Mỹ) đưa ra một báo cáo sai lệch về một cuộc tấn công toàn diện của Liên Xô vào Mỹ. Mỹ đã thực hiện các biện pháp trả đũa khẩn cấp trước khi dữ liệu từ các hệ thống radar cảnh báo sớm cho thấy rằng không có cuộc tấn công nào được tiến hành (McLean và Stewart 1979). Ngày 26 tháng 9 năm 1983, hệ thống cảnh báo sớm hạt nhân Oko của Liên Xô bị trục trặc đã báo cáo về một cuộc tấn công tên lửa của Mỹ sắp tới. Báo cáo được xác định chính xác là một báo động sai bởi sĩ quan chịu trách nhiệm tại trung tâm chỉ huy, Stanislav Petrov: một quyết định đã được ghi nhận trong việc ngăn chặn chiến tranh nhiệt hạch (Lebedev 2004). Dường như một cuộc chiến tranh có thể đã thất bại trong việc gây ra sự tuyệt chủng của con người, ngay cả khi đã diễn ra với kho vũ khí kết hợp của tất cả các cường quốc hạt nhân ở đỉnh cao của Chiến tranh Lạnh, mặc dù nó sẽ hủy hoại nền văn minh và gây ra chết chóc và đau khổ không thể hình dung được (Gaddis 1982; Parrington 1997). Nhưng các kho dự trữ lớn hơn có thể được tích lũy trong các cuộc đua vũ trang tương lai, hoặc thậm chí các vũ khí chết chóc hơn có

thể được phát minh, hoặc các mô hình về tác động của chiến tranh hạt nhân (đặc biệt là mức độ nghiêm trọng của mùa đông hạt nhân) mà chúng ta đưa ra có thể sai.

13. Cách tiếp cận này có thể phù hợp với phạm trù phương pháp kiểm soát dựa trên quy tắc đặc tả trực tiếp.

14. Tình huống về cơ bản là giống nhau nếu tiêu chí giải pháp chỉ định một biện pháp tốt chứ không phải một giới hạn cho thứ được coi là giải pháp.

15. Một người ủng hộ cách tiếp cận AI tiên tri có thể khẳng định rằng ít nhất có khả năng người dùng sẽ phát hiện ra lỗi hổng trong giải pháp được trao giải và nhận ra rằng nó không phù hợp với ý định của họ ngay cả khi đáp ứng các tiêu chí thành công được chỉ định chính thức. Khả năng bắt lỗi ở giai đoạn này sẽ phụ thuộc vào nhiều yếu tố khác nhau, bao gồm mức độ dễ hiểu đối với con người của các kết quả đầu ra của AI tiên tri và mức độ khoan dung trong việc lựa chọn các tính năng của kết quả tiềm năng để cho người dùng biết.

Ngoài ra, thay vì dựa vào chính AI tiên tri trong việc thực hiện các chức năng này, người ta có thể cố gắng xây dựng một công cụ riêng có thể kiểm tra các phát ngôn của AI tiên tri và chỉ cho chúng ta một cách hữu ích những gì sẽ xảy ra nếu hành động theo chúng. Nhưng để làm được điều này một cách tổng quát, chúng ta cần có một AI tiên tri siêu phàm khác mà sau đó chúng ta phải tin tưởng; vì vậy, vấn đề độ tin cậy sẽ không được giải quyết mà chỉ thay thế. Người ta có thể tìm cách tăng cường tính an toàn thông qua việc sử dụng nhiều AI tiên tri để thực hiện đánh giá ngang hàng, nhưng điều này không đảm bảo trong trường hợp tất cả các AI tiên tri thất bại theo cách

tương tự, như điều có thể xảy ra nếu, chẳng hạn, tất cả đều có cùng đặc điểm kỹ thuật chính thức của thứ được coi là một giải pháp thỏa đáng.

16. Với công suất tính toán đủ lớn (hữu hạn nhưng đáng tin cậy về mặt vật lý), có lẽ chúng ta sẽ đạt được sự siêu trí tuệ tổng quát với các thuật toán hiện có. (Ví dụ, so sánh với hệ thống AIXItl; Hutter [2001].) Nhưng ngay cả việc định luật Moore tiếp tục đúng trong 100 năm nữa cũng không đủ để đạt được mức độ công suất tính toán cần thiết để đạt được điều này.

17. Bird và Layzell (2002) và Thompson (1997); Xem thêm Yaeger (1994, 13-14).

18. Williams (1966).

19. Leigh (2010).

20. Ví dụ này được mượn từ Yudkowsky (2011).

21. Wade (1976). Các thí nghiệm máy tính cũng đã được thực hiện với tiến hóa mô phỏng được thiết kế giống với các khía cạnh của tiến hóa sinh học và một lần nữa cho ra kết quả kỳ lạ (Ví dụ, xem Yaeger [1994]).

CHƯƠNG 11: NHỮNG KỊCH BẢN ĐA CỰC

1. Không phải vì đây nhất thiết là loại kịch bản có khả năng nhất hoặc đáng mong đợi nhất, mà bởi đây là loại kịch bản dễ phân tích nhất với bộ công cụ của lý thuyết kinh tế tiêu chuẩn, và do đó, là điểm khởi đầu thuận tiện cho phần thảo luận này.

2. Hội đồng Ngựa Mỹ (2005). Xem thêm Salem và Rowan (2001).

3. Acemoglu (2003); Mankiw (2009); Zuleta (2008).

4. Fredriksen (2012, 8); Salverda và cộng sự. (2009, 133).

5. Nó cũng rất cần thiết cho ít nhất một phần số vốn được đầu tư vào tài sản tăng theo xu hướng chung. Một danh mục tài sản đa dạng, chẳng hạn như cổ phiếu trong quỹ theo dõi chỉ số, sẽ giúp gia tăng cơ hội không bị hoàn toàn bỏ lỡ.

6. Nhiều hệ thống phúc lợi châu Âu chưa được hoàn thành, có nghĩa là lương hưu được trả từ những người lao động hiện tại đang đóng góp và thuế thay vì từ một nguồn vốn tiết kiệm. Các chương trình như vậy sẽ không tự động đáp ứng yêu cầu (trong trường hợp thất nghiệp hàng loạt đột ngột, các khoản thu dùng để chi trả phúc lợi có thể cạn kiệt). Tuy nhiên, các chính phủ có thể chọn bù đắp sự thiếu hụt từ các nguồn khác.

7. Hội đồng Ngựa Mỹ (2005).

8. Cung cấp cho 7 tỷ người lương hưu hằng năm là 90.000 đô la/người sẽ tiêu tốn 630 nghìn tỷ Đô la/năm, gấp 10 lần GDP hiện tại của thế giới. Trong 100 năm qua, GDP thế giới đã tăng khoảng 19 lần từ khoảng hai nghìn tỷ đô la năm 1900 lên 37 nghìn tỷ đô la năm 2000 (theo tỷ giá Đô la quốc tế 1990) theo Maddison (2007). Vì vậy, nếu tốc độ tăng trưởng này tiếp diễn trong 200 năm tới, trong khi dân số không đổi, thì việc cung cấp cho mọi người khoản lương hưu 90.000 đô la/người/năm sẽ tiêu tốn khoảng 3% GDP thế giới. Một sự bùng nổ trí tuệ có thể làm cho lượng tăng trưởng này xảy ra trong một khoảng thời gian ngắn hơn nhiều. Xem thêm Hanson (1998a, 1998b, 2008).

9. Và có lẽ gấp hàng triệu lần trong 70.000 năm qua nếu có một đợt thắt cổ chai quần thể (population bottleneck) nghiêm trọng vào thời điểm đó, như đã được suy đoán. Xem thêm dữ liệu trong Kremer (1993) và Huff và cộng sự (2010).

10. Cochran và Harpending (2009). Xem thêm Clark (2007) và, xem bài phê bình trong Allen (2008).

11. Kremer (1993).

12. Basten và cộng sự. (2013). Các kịch bản xuất hiện sự gia tăng liên tục cũng có thể xảy ra. Nhìn chung, sự không chắc chắn của các dự đoán như vậy tăng lên vượt xa con số một hoặc hai thế hệ trong tương lai.

13. Trên toàn cầu, tổng mức sinh thay thế là 2,33 con/phụ nữ vào năm 2003.

Con số này xuất phát từ thực tế là phải mất hai con trên một phụ nữ để thay thế cha mẹ, cộng với “một phần ba đứa trẻ” để bù đắp (1) xác suất sinh con trai cao hơn và (2) tử vong sớm trước khi hết độ tuổi sinh đẻ. Đối với các quốc gia phát triển, con số này nhỏ hơn, khoảng 2,1, vì tỷ lệ tử vong thấp hơn. (Xem Espenshade và cộng sự [2003, Giới thiệu, Bảng 1, 580].) Dân số ở hầu hết các nước phát triển sẽ giảm nếu không tính nhập cư. Một vài ví dụ đáng chú ý của các quốc gia có tỷ lệ sinh thay thế phụ là: Singapore ở mức 0,79 (thấp nhất thế giới), Nhật Bản 1,39, Cộng hòa Nhân dân Trung Hoa 1,55, Liên minh Châu Âu 1,58, Nga 1,61, Brazil 1,81, Iran 1,86, Việt Nam 1,87 và Vương quốc Anh 1,90. Ngay cả dân số Mỹ cũng có thể sẽ giảm nhẹ với tỷ lệ sinh là 2,05. (Xem CIA [2013].)

14. Sự đầy đủ về thời gian có thể xảy ra trong hàng tỷ năm kể từ thời điểm hiện tại.

15. Carl Shulman chỉ ra rằng nếu con người sinh học sống theo tuổi thọ tự nhiên của mình bên cạnh nền kinh tế kỹ thuật số, họ cần phải thừa nhận rằng không chỉ trật tự chính trị trong lĩnh vực kỹ thuật số sẽ bảo vệ lợi ích của con người mà nó sẽ vẫn như vậy trong

khoảng thời gian rất dài (Shulman 2012). Ví dụ, nếu các sự kiện trong lĩnh vực kỹ thuật số diễn ra nhanh hơn hàng nghìn lần so với bên ngoài, thì một con người sinh học sẽ phải dựa vào việc cơ thể kỹ thuật số duy trì ổn định chính trị trong 50.000 năm thay đổi và khuấy động nội tại. Tuy nhiên, nếu thế giới chính trị kỹ thuật số giống như thế giới của chúng ta, sẽ có rất nhiều cuộc cách mạng, chiến tranh và những biến động lớn trong những thiên niên kỷ đó có thể gây bất lợi cho con người ở bên ngoài. Ngay cả rủi ro 0,01% mỗi năm của một cuộc chiến tranh nhiệt hạch toàn cầu hoặc thảm họa tương tự cũng sẽ gây ra tổn thất gần như chắc chắn cho con người sống trong một cuộc sống sinh học chậm chạp. Để khắc phục vấn đề này, sẽ cần một trật tự ổn định hơn trong lĩnh vực kỹ thuật số: có lẽ là một thể đơn nhất dần dần cải thiện sự ổn định của chính nó.

16. Người ta có thể nghĩ rằng ngay cả khi máy móc có hiệu quả cao hơn nhiều so với con người, vẫn sẽ có một mức lương nào đó mà việc sử dụng công nhân là con người mang lại lợi nhuận, chẳng hạn ở mức 1 cent/giờ. Nếu đây là nguồn thu nhập duy nhất cho con người, nhân loại sẽ tuyệt chủng vì loài người không thể tồn tại chỉ bằng 1 cent/giờ. Nhưng con người cũng có được thu nhập từ vốn. Bây giờ, nếu chúng ta giả định rằng dân số tăng lên cho đến khi tổng thu nhập ở mức đủ sống, người ta có thể nghĩ rằng đây sẽ là một trạng thái mà con người sẽ làm việc chăm chỉ. Giả sử, thu nhập ở mức đủ sống là 1 Đô la/ngày. Sau đó, có vẻ như, dân số sẽ tăng lên cho đến khi vốn trên đầu người chỉ mang lại thu nhập 90 cent/ngày, và mọi người sẽ phải bổ sung 10 giờ lao động nặng nhọc để kiếm 10 cent còn lại. Tuy nhiên, tình hình không cần phải như vậy, bởi vì thu nhập ở mức đủ sống phụ thuộc vào khối lượng công việc được thực

hiện: những người làm việc chăm chỉ hơn đốt cháy nhiều calo hơn. Giả sử, mỗi giờ làm việc làm tăng chi phí thực phẩm thêm 2 cent. Từ đó, chúng ta có một mô hình mà trong đó, con người đứng yên ở một điểm cân bằng.

17. Có thể nghĩ rằng một cuộc họp kín như thế này sẽ không thể bỏ phiếu hoặc bảo vệ các quyền lợi của nó. Nhưng những người sống ở đây có thể trao giấy ủy quyền cho các ủy ban AI để quản lý công việc của họ và đại diện cho lợi ích chính trị của họ. (Phần này của cuộc thảo luận được đặt ra dựa trên giả định rằng các quyền sở hữu được tôn trọng.)

18. Không rõ từ ngữ tốt nhất là gì. “Giết” có thể gợi ý sự tàn bạo tích cực hơn mức được đảm bảo. “Kết thúc” có thể mang quá nhiều nét uyển ngữ. Sự phức tạp nằm ở chỗ có hai sự kiện có thể tách bạch: ngừng chạy một quy trình và xóa mẫu thông tin. Cái chết của con người thường liên quan đến cả hai sự kiện, nhưng đối với một bản giả lập, chúng có thể là hai sự kiện riêng. Việc một chương trình tạm thời ngừng chạy có thể không có hậu quả lớn hơn việc một người đi ngủ: nhưng ngừng chạy vĩnh viễn có thể tương đương với hôn mê vĩnh viễn. Tuy nhiên, sự phức tạp phát sinh xa hơn từ việc các bản giả lập có thể được sao chép và chạy ở các tốc độ khác nhau: con người không trải qua những thứ tương tự. (So sánh với Bostrom [2006b]; Bostrom và Yudkowsky [2015].)

19. Sẽ có sự đánh đổi giữa tổng công suất tính toán song song và tốc độ tính toán, vì tốc độ tính toán cao nhất sẽ chỉ có thể đạt được khi giảm công suất. Điều này sẽ đặc biệt đúng sau khi một người bước vào kỷ nguyên điện toán đảo ngược.

20. Một bản giả lập có thể được kiểm tra bằng cách đưa nó vào một sự cám dỗ. Bằng cách liên tục kiểm tra cách bản giả lập bắt đầu từ một trạng thái chuẩn bị nhất định phản ứng với các chuỗi kích thích khác nhau, người ta có thể có được tự tin cao về độ tin cậy của bản giả lập đó. Nhưng càng về sau, trạng thái tinh thần càng được cho phép phát triển ra khỏi điểm bắt đầu được xác nhận của nó, thì càng ít chắc chắn rằng nó sẽ vẫn đáng tin cậy.

21. Một vài bản giả lập có thể giống với các bản giả lập cùng nhóm (tất cả các bản sao và biến thể có nguồn gốc từ cùng một mẫu) hơn so với bất kỳ bản khởi tạo cụ thể nào. Một bản giả lập như vậy có thể không coi sự kết thúc của mình là một sự kiện chết chóc, nếu nó biết rằng các thành viên khác trong nhóm sẽ sống sót. Các bản giả lập có thể biết rằng chúng sẽ được hoàn nguyên về trạng thái lưu trữ cụ thể vào cuối ngày và mất đi ký ức của ngày hôm đó, nhưng hãy bỏ qua điều này giống như một người tham gia bữa tiệc biết rằng cô ta sẽ thức dậy vào sáng hôm sau mà không có bất kỳ hồi ức nào về đêm trước đó: coi đây là mất trí nhớ, không phải là chết.

22. Một đánh giá đạo đức cũng có thể tính đến nhiều yếu tố khác. Ngay cả khi toàn bộ công nhân liên tục hài lòng với tình trạng của họ, kết quả vẫn có thể bị phản đối sâu sắc về đạo đức về các vấn đề khác, mặc dù những căn cứ đó là vấn đề tranh chấp giữa các lý thuyết đạo đức đối lập. Nhưng bất kỳ đánh giá hợp lý nào cũng sẽ coi hạnh phúc chủ quan là một yếu tố quan trọng. Xem thêm Bostrom và Yud-kowsky (2015).

23. Khảo sát Giá trị Thế giới (2008).

24. Helliwell và Sachs (2012).

25. So sánh với Bostrom (2004). Xem thêm Chislenko (1996) và Moravec (1988).

26. Thật khó để nói liệu các cấu trúc xử lý thông tin xuất hiện trong loại kịch bản này có ý thức hay không (theo nghĩa là có qualia, trải nghiệm mang tính hiện tượng). Điều này khó một phần là do sự thiếu hiểu trải nghiệm của chúng ta về việc những thực thể nhận thức nào sẽ phát sinh và một phần là sự thiếu hiểu biết về triết học của chúng ta về việc loại cấu trúc nào là có ý thức. Người ta có thể cố gắng điều chỉnh lại câu hỏi, và thay vì hỏi liệu các thực thể trong tương lai có ý thức hay không, người ta có thể hỏi liệu các thực thể trong tương lai có vị thế đạo đức hay không; hoặc liệu chúng có phải là các cá thể mà chúng ta coi sự “khỏe mạnh” của họ là ưu tiên hay không. Nhưng những câu hỏi này có thể không dễ trả lời hơn câu hỏi về ý thức, thực tế, họ có thể yêu cầu một câu trả lời cho câu hỏi ý thức vì vị thế đạo đức hoặc ưu tiên của chúng ta phụ thuộc vào việc thực thể trong câu hỏi có thể chủ quan trải nghiệm tình trạng của nó hay không.

27. Đối với một lập luận rằng cả lịch sử địa chất và con người đều hướng đến một sự phức tạp lớn hơn, xem Wright (2001). Đối với một lập luận đối nghịch (bị chỉ trích trong Chương 9 cuốn sách của Wright), xem Gould (1990). Xem thêm lập luận rằng chúng ta đang chứng kiến một xu hướng thuyên giảm bạo lực và sự tàn bạo mạnh mẽ và lâu dài trong Pinker (2011).

28. Xem thêm về lý thuyết chọn lọc quan sát trong Bostrom (2002a).

29. Bostrom (2008a). Một nghiên cứu tiến hành cẩn thận hơn nhiều về các chi tiết của lịch sử tiến hóa của chúng ta sẽ là cần thiết

để tránh hiệu ứng chọn lọc. Ví dụ, xem Carter (1983, 1993); Hanson (1998d); Ćirković và cộng sự (2010).

30. Kansa (2003).

31. Ví dụ: Zahavi and Zahavi (1997).

32. Xem Miller (2000).

33. Kansa (2003). Xem trường hợp có tính gợi mở hơn trong Frank (1999).

34. Cách tốt nhất để đo lường mức độ hội nhập chính trị toàn cầu là không rõ ràng. Một viễn cảnh có thể sẽ là trong khi một bộ tộc hái lượm săn bắt có thể đã hợp nhất 100 cá nhân thành một thực thể ra quyết định thì các thực thể chính trị lớn nhất hiện nay chứa hơn một tỷ cá nhân. Điều này sẽ lên tới mức chênh lệch bảy cấp khuếch đại (107 lần), chỉ cần nhân thêm 10 lần trước khi toàn bộ dân số thế giới được đưa vào một thực thể chính trị duy nhất. Tuy nhiên, vào thời điểm bộ tộc này có quy mô hội nhập lớn nhất, dân số thế giới nhỏ hơn nhiều. Bộ tộc có thể chứa tới một phần nghìn dân số thời điểm đó. Điều này sẽ làm cho sự gia tăng quy mô hội nhập chính trị chỉ là khoảng 100 lần. Nhìn vào tỷ lệ dân số thế giới được hội nhập chính trị, thay vì số lượng tuyệt đối, xem ra là phù hợp trong bối cảnh hiện tại (đặc biệt là việc chuyển đổi sang trí tuệ máy có thể gây ra một sự “bùng nổ dân số” của các bản mô phỏng hoặc các kiểu tâm trí kỹ thuật số khác). Nhưng cũng đã có những phát triển trong các tổ chức toàn cầu và mạng lưới phối hợp bên ngoài các cấu trúc nhà nước chính thức, và chúng cũng cần được tính đến.

35. Một trong những lý do cho rằng cuộc cách mạng trí tuệ máy đầu tiên sẽ xảy ra nhanh chóng (sự tồn tại có thể của một mái hiên phần cứng) không đúng ở đây. Tuy nhiên, có thể có các nguồn tăng

nhANH khác, chẳng hạn như một bước đột phá mạnh mẽ trong phần mềm liên quan đến việc chuyển đổi từ giả lập sang trí tuệ máy tổng hợp thuần túy.

36. Shulman (2010b).

37. Làm thế nào cân bằng các điểm lợi và hại có thể phụ thuộc vào loại công việc mà siêu sinh vật này đang cố gắng thực hiện, và mẫu giả lập có năng lực tổng thể cao nhất tồn tại sẽ có năng lực tổng thể ra sao. Và các tổ chức lớn hiện nay cần sở hữu bộ máy con người phong phú một phần vì các cá nhân đa tài là vô cùng hiếm gặp.

38. Tất nhiên việc tạo nhiều bản sao của một tác tử phần mềm là rất dễ dàng.

Nhưng lưu ý rằng việc sao chép nhìn chung là không đủ để đảm bảo rằng các bản sao có cùng mục tiêu tối thượng. Để hai tác tử có cùng mục tiêu tối thượng, các mục tiêu phải trùng khớp nhau về các yếu tố mang tính chỉ biệt. Nếu Bob ích kỷ, một bản sao của Bob cũng sẽ ích kỷ. Tuy nhiên, mục tiêu của họ không trùng khớp: Bob quan tâm đến Bob trong khi Bob-copy quan tâm đến Bob-copy.

39. Shulman (2010b, 6).

40. Điều này có thể khả thi hơn đối với người sinh học và giả lập hoàn chỉnh não bộ so với trí tuệ nhân tạo tự do, có thể được xây dựng để có các bộ phận hoặc động lực chức năng ẩn rất khó khám phá. Mặt khác, các AI được chế tạo đặc biệt rõ ràng sẽ cho phép kiểm tra và xác minh kỹ lưỡng hơn so với các kiến trúc giống não bộ. Áp lực xã hội có thể khuyến khích các AI phơi bày mã nguồn của mình và tự sửa đổi để hiển thị bản thân một cách rõ ràng, đặc biệt

nếu rõ ràng là điều kiện tiên quyết để được tin tưởng và do đó được trao cơ hội tham gia vào các giao dịch có lợi. So sánh với Hall (2007).

41. Một số vấn đề khác có vẻ tương đối nhỏ, đặc biệt là trong trường hợp các khoản đóng góp là rất lớn (tương đương với các khoản đóng góp vào những thất bại chính trong việc phối hợp toàn cầu), bao gồm chi phí tìm kiếm các chính sách vì lợi ích chung, và khả năng một số tác tử có thể có một ưu tiên cơ bản cho “sự tự chủ” của người dùng ở một dạng thức sẽ được giảm bằng cách tham gia vào các hiệp ước toàn cầu có các cơ chế giám sát và thực thi kèm theo.

42. Một AI có thể đạt được điều này bằng cách tự sửa đổi một cách thích hợp và sau đó cho người quan sát quyền truy cập chỉ đọc vào mã nguồn của nó. Một trí tuệ máy với kiến trúc mờ hơn (như một bản giả lập) có thể đạt được nó bằng cách áp dụng công khai cho chính mình một phương pháp lựa chọn động lực nào đó. Ngoài ra, một cơ quan cưỡng chế bên ngoài, chẳng hạn như lực lượng cảnh sát siêu sinh vật, có lẽ có thể được sử dụng không chỉ để thực thi một hiệp ước đạt được giữa các bên khác nhau, mà còn được sử dụng nội bộ bởi một bên duy nhất cam kết thực hiện một hành động cụ thể.

43. Chọn lọc tiến hóa có thể ủng hộ những người không biết đến mối đe dọa và thậm chí cả những người tỏ ra gồng mình đến mức thà chiến đấu đến chết còn hơn chịu thất bại nhỏ nhất. Khuynh hướng này có thể mang lại cho người sở hữu nó nhiều lợi ích lớn. (Bất kỳ phần thưởng mang tính công cụ nào như vậy tất nhiên không cần đóng vai trò gì trong động lực có ý thức của tác tử: anh ta có thể coi trọng công lý hoặc danh dự như là mục đích của bản thân.)

44. Tuy nhiên, để có một “bản án” dứt khoát về những vấn đề này, ta cần những phân tích sâu hơn. Có nhiều biến thể tiềm tàng khác mà chúng ta không thể khám phá ở đây.

CHƯƠNG 12: THỤ ĐẮC CÁC GIÁ TRỊ

1. Các biến thể khác nhau của ý tưởng cơ bản này có thể được giới thiệu.

Chúng ta đã thảo luận về một biến thể trong Chương 8 (là tác tử thỏa mãn vừa đủ, trái ngược với tối đa hóa), và trong chương tiếp theo, chúng ta đề cập ngắn gọn về vấn đề các lý thuyết quyết định thay thế. Tuy nhiên, những vấn đề như vậy không phải là thiết yếu đối với trọng tâm của tiểu mục này, vì vậy chúng tôi sẽ giữ mọi thứ đơn giản bằng cách tập trung vào trường hợp của một tác tử tối đa hóa độ thỏa dụng kỳ vọng.

2. Giả sử AI là có hàm thỏa dụng đáng kể. Việc xây dựng một tác tử luôn chọn một hành động tối đa hóa độ thỏa dụng kỳ vọng sẽ rất dễ dàng nếu hàm thỏa dụng của nó là, ví dụ, hàm hằng $U(w) = 0$. Mọi hành động đều tối đa hóa độ thỏa dụng kỳ vọng liên quan đến hàm thỏa dụng đó.

3. Cũng bởi chúng ta đã quên đi sự nhầm lẫn màu mè thời thơ ấu, thời điểm mà chúng ta chưa thể nhìn rõ vì não bộ chưa học cách giải thích thông tin đầu vào hình ảnh của nó.

4. Xem thêm Yudkowsky (2011) và đánh giá trong phần 5 của Muehlhauser và Helm (2012).

5. Có lẽ việc những tiến bộ trong công nghệ phần mềm cuối cùng có thể vượt qua những khó khăn này chỉ là vừa đủ dễ hiểu. Với các công cụ hiện đại, một lập trình viên duy nhất có thể tạo ra phần mềm vượt xa một nhóm các nhà phát triển lớn buộc phải viết trực tiếp

bằng mã máy. Các lập trình viên AI ngày nay có được khả năng diễn đạt (expressiveness) từ sự phổ biến của các thư viện tính toán khoa học và máy tính chất lượng cao, cho phép ai đó hack, ví dụ, một ứng dụng đếm khuôn mặt trên webcam bằng cách xâu chuỗi các thư viện mà họ không bao giờ tự viết được. Việc tích lũy phần mềm có thể tái sử dụng, được sản xuất bởi các chuyên gia nhưng những người không chuyên cũng sử dụng được, sẽ mang lại cho các lập trình viên tương lai một lợi thế về khả năng diễn đạt. Ví dụ, một lập trình viên robot trong tương lai có thể truy cập dễ dàng vào các thư viện nét mặt tiêu chuẩn, các bộ sưu tập đối tượng xây dựng văn phòng điển hình, các thư viện quỹ đạo chuyên biệt và nhiều chức năng khác hiện không sẵn có.

6. Dawkins (1995, 132). Yêu cầu ở đây không nhất thiết là định lượng đau khổ trong thế giới tự nhiên lớn hơn định lượng hạnh phúc tích cực.

7. Kích thước quần thể cần thiết có thể lớn hơn hoặc nhỏ hơn nhiều so với kích thước tồn tại ở thời kỳ của tổ tiên của chúng ta. Xem Shulman và Bostrom (2012).

8. Nếu dễ dàng có được một kết quả tương đương mà không gây tổn hại cho số lượng lớn người vô tội, thì thực hiện việc này xem ra tốt hơn về đạo đức. Tuy nhiên, nếu những cá thể người kỹ thuật số được tạo ra và chịu thiệt hại một cách bất công, việc bù đắp cho sự đau khổ của họ bằng cách lưu họ vào file và (khi tương lai của loài người được bảo đảm) chạy lại trong những điều kiện thuận lợi hơn sẽ là khả thi. Sự bù đắp như vậy có thể so sánh với các quan niệm tôn giáo về thế giới bên kia khi ngành thần học cố gắng giải quyết các vấn đề về sự tà ác.

9. Một trong những nhân vật hàng đầu của lĩnh vực này, Richard Sutton, định nghĩa học tăng cường không phải về một phương pháp học tập mà là một vấn đề học tập: bất kỳ phương pháp nào phù hợp để giải quyết vấn đề đó đều được coi là phương pháp học tăng cường (Sutton và Barto 1998, 4). Ngược lại, các thảo luận hiện tại liên quan đến các phương pháp mà tác tử có thể được coi là lấy tối đa hóa (một số ý niệm về) phần thưởng tích lũy làm mục tiêu tối thượng. Vì một tác tử với một loại mục tiêu tối thượng nào đó rất khác có thể có kỹ năng bắt chước một tác tử tìm kiếm phần thưởng trong một loạt các tình huống, và do đó có thể rất phù hợp để giải quyết các vấn đề học tăng cường, nên có thể có các phương pháp được tính là “các phương pháp học tăng cường” theo định nghĩa của Sutton sẽ không dẫn đến hội chứng wireheading. Tuy nhiên, các nhận xét ở đây áp dụng cho hầu hết các phương pháp thực sự được sử dụng trong cộng đồng học tăng cường.

10. Ngay cả khi, bằng cách nào đó, một cơ chế giống con người có thể được thiết lập trong trí tuệ máy giống con người, những mục tiêu tối thượng mà trí tuệ này đạt được không cần giống với một người được điều chỉnh tốt, trừ khi môi trường nuôi dưỡng cho đứa trẻ kỹ thuật số này cũng sát với một đứa trẻ bình thường: một thứ gì đó khó sắp xếp. Và ngay cả với một môi trường nuôi dưỡng giống như con người, chúng ta cũng không đảm bảo được một kết quả khả quan, vì ngay cả một sự khác biệt nhỏ trong khuynh hướng bẩm sinh cũng có thể dẫn đến những phản ứng rất khác nhau đối với một sự kiện trong cuộc sống. Tuy nhiên, có thể tạo ra một cơ chế bồi đắp giá trị đáng tin cậy hơn cho các “bộ óc” giống con người trong tương lai

(có thể sử dụng các loại thuốc mới hoặc cấy ghép não, hoặc các phiên bản kỹ thuật số tương đương của chúng).

11. Người ta có thể tự hỏi tại sao con người chúng ta xem ra không cố gắng vô hiệu hóa cơ chế dẫn đến việc thụ đắc những giá trị cuối cùng mới. Một số yếu tố có thể đang tồn tại. Đầu tiên, hệ thống động lực của con người được mô tả sơ sài như một thuật toán tối đa hóa độ thỏa dụng tính toán lạnh lùng. Thứ hai, chúng ta có thể không có bất kỳ phương tiện thuận tiện nào để thay đổi cách thức thụ đắc các giá trị. Thứ ba, chúng ta có thể có những lý do công cụ (phát sinh, ví dụ, từ nhu cầu phát tín hiệu xã hội) đôi khi có được các giá trị cuối cùng mới – các giá trị công cụ có thể không hữu ích nếu một phần tâm trí của chúng ta hiện ra rõ rệt trong mắt những người khác, hoặc nếu sự phức tạp về nhận thức của việc giả vờ có được một tập hợp các giá trị cuối cùng khác với những gì chúng ta có trong thực tế gây ra quá nhiều mệt mỏi. Thứ tư, có những trường hợp chúng ta chủ động chống lại các xu hướng tạo ra thay đổi trong các giá trị cuối cùng của mình, ví dụ như khi chúng ta tìm cách chống lại những ảnh hưởng tha hóa từ người xấu xung quanh. Thứ năm, có một khả năng thú vị là chúng ta đặt một số giá trị cuối cùng vào việc trở thành loại tác tử có thể thụ đắc các giá trị cuối cùng mới theo cách thông thường của con người.

12. Hoặc người ta có thể cố gắng thiết kế hệ thống động lực để AI không quan tâm đến sự thay thế đó; xem Armstrong (2010).

13. Ở đây chúng ta sẽ tận dụng một vài giải thích của Daniel Dewey (2011).

Các ý tưởng nền tảng khác đóng góp cho khung này đã được phát triển bởi Marcus Hutter (2005), Shane Legg (2008), Eliezer

Yudkowsky (2001), Nick Hay (2005), Moshe Looks và Peter de Blanc.

14. Để tránh các biến thể không cần thiết, chúng ta giới hạn sự chú ý của mình đến các tác tử kiểu quyết định luận không coi nhẹ phần thưởng trong tương lai.

15. Về mặt toán học, một hành vi của tác tử có thể được hình thức hóa như là một hàm tác tử (agent function), ánh xạ mỗi lịch sử tương tác khả thể vào một hành động. Ngoại trừ các tác tử rất đơn giản, ta không thể biểu diễn rõ ràng chức năng của tác tử như một bảng tra cứu. Thay vào đó, tác tử được cung cấp một số cách tính toán để thực hiện hành động. Vì có nhiều cách tính toán cùng một hàm tác tử, điều này dẫn đến sự phân chia rõ ràng hơn của một tác tử: một chương trình tác tử. Một chương trình tác tử là một chương trình hoặc thuật toán cụ thể tính toán một hành động cho bất kỳ lịch sử tương tác nào. Mặc dù việc nghĩ về một chương trình tác tử tương tác với một số môi trường được chỉ định chính thức thường rất thuận tiện và hữu ích về mặt toán học, điều quan trọng cần nhớ là đây là một sự lý tưởng hóa. Tác tử thực sự được khởi tạo bằng phương pháp vật lý. Điều này không chỉ có nghĩa là tác tử tương tác với môi trường thông qua các cơ quan thụ cảm và cơ quan chấp hành của nó, mà còn là việc “não bộ” hay bộ điều khiển của tác tử chính là một phần của thực tại vật lý. Do đó, hoạt động của nó về nguyên tắc có thể bị ảnh hưởng bởi các can thiệp vật lý từ bên ngoài (và không chỉ bằng cách nhận được các đối tượng tri giác từ các cơ quan thụ cảm của nó). Do đó, tại một số điểm, cần phải xem một tác tử là một bản thực thi tác tử (agent implementation). Bản thực thi tác tử là một cấu trúc vật lý sẽ thực thi một hàm tác tử trong trường hợp

không có sự can thiệp từ môi trường. (Theo định nghĩa của Dewey [2011].)

16. Dewey đề xuất khái niệm tối ưu sau đây cho một tác tử học giá trị:

Ở đây, P_1 và P_2 là hai hàm xác suất. Tổng thứ hai nằm trong một số lớp hàm thỏa dụng phù hợp với lịch sử tương tác khả thể. Trong phiên bản được trình bày ở đây, chúng tôi đã đưa ra một số phần phụ thuộc rõ ràng cũng như tận dụng khái niệm đơn giản hóa về các thế giới khả thể.

17. Cần lưu ý rằng tập hợp U các hàm thỏa dụng phải thỏa mãn việc các giá trị về độ thỏa dụng có thể được so sánh và tính trung bình. Nói chung, việc này không đơn giản, và cách biểu diễn các lý thuyết đạo đức khác nhau về cái “tốt” qua các hàm thỏa dụng chính là không rõ ràng. Ví dụ, xem MacAskill 2010.

18. Hay nói chung hơn, vì V có thể không giống như ngụ ý trực tiếp cho bất kỳ cặp nào của một thế giới khả thể và một hàm thỏa dụng (w, U) cho dù mệnh đề $V(U)$ có đúng trong w hay không, điều cần làm là cung cấp cho AI một phép biểu diễn đầy đủ của phân phối xác suất có điều kiện $P(V(U)|w)$.

19. Đầu tiên hãy xét tập hợp Y , tập hợp hành động khả thể cho một tác tử.

Vấn đề ở đây là chính xác điều gì nên được tính là một hành động: các lệnh vận động (motor command) cơ bản (ví dụ: “gửi một xung điện dọc theo kênh đầu ra #00101100”), hay các hành động cấp cao hơn (ví dụ: “giữ camera ở giữa khuôn mặt”)? Vì đang cố gắng phát triển một khái niệm tối ưu hơn là một kế hoạch triển khai thực tế, chúng ta có thể coi Y là các lệnh vận động cơ bản (và vì tập hợp các

lệnh vận động có thể thay đổi theo thời gian, chúng ta có thể cần lập chỉ mục Y theo thời gian). Tuy nhiên, để tiến tới thực hiện, có lẽ cần phải đưa vào một số loại quy trình lập kế hoạch phân cấp, và sau đó người ta có thể cần xem xét cách áp dụng công thức cho một số loại hành động cấp cao hơn. Một vấn đề khác là làm thế nào để phân tích các hành động nội bộ (chẳng hạn như viết các chuỗi vào bộ nhớ làm việc). Vì các hành động nội bộ có thể có những hệ quả quan trọng, chúng ta sẽ muốn Y bao gồm các hành động nội bộ cơ bản như vậy, cũng như các lệnh vận động. Nhưng có những giới hạn về phạm vi mà ta có thể đi theo hướng này: việc tính toán độ thỏa dụng kỳ vọng của bất kỳ hành động nào trong Y đòi hỏi nhiều phép tính toán, và nếu mỗi phép tính đó cũng được coi là một hành động trong Y cần được đánh giá theo AI-VL, chúng ta sẽ phải đối mặt với một hồi quy vô hạn khiến nó không thể bắt đầu được. Để tránh hồi quy vô hạn, người ta phải hạn chế mọi nỗ lực rõ ràng nhằm ước tính độ thỏa dụng kỳ vọng ở một số lượng hạn chế các khả năng hành động đáng kể. Hệ thống sau đó sẽ cần một số quy trình tự nghiệm xác định một số khả năng hành động đáng kể để xem xét thêm. (Cuối cùng, hệ thống cũng có thể đưa ra các quyết định rõ ràng liên quan đến một số hành động khả thể nhằm sửa đổi quy trình tự nghiệm này, các hành động có thể bị “dán cờ” (flagged) để được chính quá trình này chú ý đến một cách cụ thể; vì vậy về lâu dài, hệ thống có thể trở nên ngày càng hiệu quả trong việc tiệm cận lý tưởng được xác định bởi AI-VL.)

Kế tiếp, hãy xem xét W , đó là một lớp các thế giới khả thể. Một khó khăn ở đây là chỉ rõ W sao cho nó đủ bao chứa. Việc không bao chứa một số w có liên quan trong W có thể khiến AI không có khả năng biểu diễn một tình huống thực sự xảy ra, dẫn đến việc AI đưa

ra quyết định không chính xác. Giả sử, chúng ta sử dụng một lý thuyết bản thể học nào đó để xác định cấu trúc của W . Chẳng hạn, chúng ta đưa vào W tất cả các thế giới khả thể bao gồm một lựa chọn không thời gian đa chiều nhất định nhất định được tạo ra bởi các hạt cơ bản trong mô hình chuẩn của vật lý hạt. Điều này có thể làm biến dạng nhận thức luận của AI nếu mô hình chuẩn không đầy đủ hoặc không chính xác. Có thể cố gắng sử dụng một tập hợp W lớn hơn để có nhiều khả năng hơn; nhưng ngay cả khi có thể đảm bảo rằng mọi vũ trụ vật lý có thể được bao chứa, người ta vẫn có thể lo lắng rằng một số khả năng khác sẽ bị loại bỏ. Ví dụ, khả năng xuất hiện các thế giới nhị nguyên khả thể mà trong đó, những sự thật về ý thức không xuất phát từ những sự thật về vật lý thì sao? Những sự thật mang tính chỉ biệt thì thế nào? Thế còn những loại thực thể khác mà con người chúng ta có thể bỏ qua nhưng điều đó có thể trở nên quan trọng để khiến mọi thứ diễn ra tốt đẹp như thế nào? Một số người có niềm tin mạnh mẽ rằng một số lý thuyết bản thể học cụ thể là chính xác. (Trong số những người viết về tương lai của AI, niềm tin vào một bản thể học duy vật, trong đó tinh thần xuất phát từ vật chất, thường được coi là điều hiển nhiên.) Tuy nhiên, một khoảnh khắc suy nghĩ về lịch sử các ý tưởng sẽ giúp chúng ta nhận ra rằng có một khả năng đáng kể của việc bản thể học mà chúng ta yêu thích là sai. Nếu các nhà khoa học thế kỷ 19 đã thử định đưa ra một nghĩa lấy cảm hứng từ vật lý của W , họ có thể đã bỏ qua khả năng về một không thời gian phi Euclid, một lý thuyết lượng tử “diễn giải đa thế giới”, một giả thuyết đa vũ trụ: các khả năng hiện có xác suất đạt được đáng kể trong thế giới thực. Có thể có những khả năng khác mà chúng ta trong thế hệ hiện tại cũng bị lãng quên giống thế. (Mặt

khác, nếu W quá lớn, sẽ nảy sinh những khó khăn về công nghệ liên quan đến việc phải gán giới hạn cho các tập hợp vô hạn.) Điều lý tưởng có thể là chúng ta sắp xếp những thứ mà AI có thể sử dụng một loại bản thể học “kết thúc mở” (open-ended) nào đó, thứ mà chính AI sau đó có thể mở rộng bằng cách sử dụng các nguyên tắc tương tự mà chúng ta sẽ sử dụng khi quyết định có công nhận một loại khả năng siêu hình mới hay không.

Xét $P(w|E_y)$. Việc chỉ rõ xác suất có điều kiện này không hoàn toàn là một phần của vấn đề truyền tải giá trị. Để có trí tuệ, AI phải có một số cách để có được xác suất chính xác một cách hợp lý so với nhiều khả năng thực tế có liên quan. Một hệ thống thiếu hụt điểm số này quá xa sẽ không gây ra mối nguy hiểm khiến chúng ta phải lo ngại. Tuy nhiên, có thể có nguy cơ AI sẽ kết thúc với một nhận thức luận đủ tốt để khiến nó hoạt động hiệu quả nhưng không đủ tốt để cho phép nó suy nghĩ chính xác về một số khả năng có tầm quan trọng chuẩn mực lớn. (Vấn đề chỉ rõ $P(w|E_y)$ theo cách này liên quan đến vấn đề chỉ rõ W .) Chỉ rõ $P(w|E_y)$ cũng đòi hỏi chúng ta đối mặt với các vấn đề khác, chẳng hạn như làm thế nào để thể hiện sự không chắc chắn đối với những sự bất khả về logic.

Các vấn đề đã nói ở trên (cách thức xác định một lớp các hành động khả thể, một lớp các thế giới khả thể và phân phối khả năng kết nối bằng chứng với các lớp của thế giới khả thể) là khá chung chung: các vấn đề tương tự nảy sinh đối với một loạt các tác tử được chỉ rõ về hình thức. Chúng ta vẫn còn phải xem xét một tập hợp các vấn đề đặc biệt hơn đối với phương pháp học giá trị; cụ thể là cách xác định U , $V(U)$ và $P(V(U)|w)$. U là một lớp các hàm thỏa dụng. Có một mối liên hệ giữa U và W vì mỗi hàm thỏa dụng $U(w)$ trong U nên gán các

tiện ích cho mỗi thế giới khả thể có trong W . Nhưng U cũng cần phải rộng để chứa đủ nhiều hàm thỏa dụng đa dạng để chúng ta có thể tự tin rằng ít nhất một trong số chúng làm tốt việc biểu diễn các giá trị dự định.

Lý do viết $P(V(U)|w)$ chứ không chỉ đơn giản là $P(U|w)$ là để nhấn mạnh việc xác suất được gán cho các mệnh đề. Một hàm thỏa dụng về bản chất không phải là một mệnh đề, nhưng chúng ta có thể chuyển đổi một hàm thỏa dụng thành một mệnh đề bằng cách đưa ra một số yêu cầu về nó. Ví dụ: chúng ta có thể yêu cầu một hàm thỏa dụng cụ thể $U(.)$ mô tả các ưu tiên của một người cụ thể hoặc biểu diễn các quy định được ngụ ý bởi một lý thuyết đạo đức nào đó, hoặc là hàm thỏa dụng mà chủ thể mong muốn phải thực hiện khi đã cân nhắc kỹ mọi thứ. Do đó, “tiêu chí giá trị” $V(.)$ Có thể được hiểu là một hàm lấy đối số của nó là hàm thỏa dụng U và đưa ra giá trị của nó là một mệnh đề cho việc U thỏa mãn tiêu chí V . Một khi đã xác định một mệnh đề $V(U)$, chúng ta hy vọng có thể có được xác suất có điều kiện $P(V(U)|w)$ từ bất kỳ nguồn nào sử dụng để có được các phân phối xác suất khác trong AI. (Nếu chắc chắn rằng tất cả các sự kiện có liên quan thông thường được tính đến khi phân chia các thế giới khả thể W , thì $P(V(U)|w)$ nên bằng 0 hoặc 1 trong mỗi thế giới khả thể.) Câu hỏi vẫn là làm thế nào để xác định V . Điều này được thảo luận thêm trong sách.

20. Đây không phải là những thách thức duy nhất cho phương pháp học giá trị. Chẳng hạn, một vấn đề khác là làm thế nào để AI có niềm tin ban đầu đủ nhạy cảm, ít nhất là vào lúc nó trở nên đủ mạnh để lật đổ các lập trình viên cố gắng cải tạo nó.

21. Yudkowsky (2001).

22. Thuật ngữ này được lấy từ bóng bầu dục: “Hail Mary” là một đường chuyền rất dài được thực hiện trong tuyệt vọng. Diễn hình là khi thời gian sắp hết, rất có thể một cầu thủ sẽ bắt bóng ở gần khu vực cuối và ghi bàn.

23. Cách tiếp cận Hail Mary dựa trên ý tưởng rằng một siêu trí tuệ có thể nói rõ sở thích của nó với độ chính xác cao hơn con người chúng ta. Ví dụ, một siêu trí tuệ có thể chỉ định ưu tiên của nó là mã. Vì vậy, nếu AI của chúng ta biểu diễn các siêu trí tuệ khác dưới dạng các quá trình tính toán đang cảm nhận môi trường của chúng, thì nó sẽ có thể suy luận về cách các siêu trí tuệ ngoài hành tinh đó phản ứng với một số kích thích giả định, như một “cửa sổ” xuất hiện trong thị trường và cho chúng thấy mã nguồn của AI của chúng ta và yêu cầu chúng làm rõ hướng dẫn của họ theo một số định dạng thuận tiện được chỉ định trước. AI của chúng ta sau đó có thể đọc được các hướng dẫn tưởng tượng này (từ mô hình của chính kịch bản giả mà trong đó các siêu trí tuệ ngoài hành tinh này xuất hiện), và chúng ta sẽ xây dựng AI của mình để nó có động lực làm theo các hướng dẫn đó.

24. Một cách khác là tạo ra một máy dò “lục lọi” (trong mô hình thế giới của AI của chúng ta) để tìm (các phép biểu diễn của) các cấu trúc vật lý được tạo ra bởi một nền văn minh siêu trí tuệ. Sau đó, chúng ta có thể bỏ qua bước xác định các hàm ưu tiên của siêu trí tuệ giả thuyết, và cung cấp cho AI của chính chúng ta giá trị cuối cùng của việc cố gắng sao chép bất kỳ cấu trúc vật lý nào mà nó tin rằng các nền văn minh siêu trí tuệ có xu hướng tạo ra.

Tuy nhiên, phiên bản này có những thách thức kỹ thuật. Chẳng hạn, vì AI của chính chúng ta, ngay cả sau khi đạt được siêu trí tuệ,

có thể không biết chính xác những cấu trúc vật lý nào được các siêu trí tuệ khác xây dựng, nó có thể cần phải cố gắng ước lượng các cấu trúc đó. Để làm điều này, có vẻ như AI của chúng ta sẽ cần một thước đo tương tự để đánh giá xem vật phẩm vật lý tiệm cận một vật phẩm khác ra sao. Nhưng các số liệu tương tự dựa trên các biện pháp vật lý thô có thể không đầy đủ (chẳng hạn, nó không đủ tốt để đánh giá rằng một bộ não giống với phô mai Camembert hơn là một máy tính chạy giả lập).

Một cách tiếp cận khả thi hơn có thể là tìm kiếm các “đèn hiệu” (beacon): các thông điệp về các hàm thỏa dụng được mã hóa theo một số định dạng đơn giản phù hợp. Chúng ta sẽ xây dựng AI của mình để nó muốn theo dõi bất cứ thông điệp nào như vậy về các hàm thỏa dụng mà nó giả thuyết rằng có thể tồn tại ngoài vũ trụ; và chúng ta hy vọng rằng các AI ngoài Trái đất thân thiện sẽ tạo ra nhiều loại đèn hiệu mà chúng (với siêu trí tuệ) cho rằng các nền văn minh đơn giản như chúng ta có khả năng xây dựng AI để tìm kiếm.

25. Nếu mọi nền văn minh đều cố gắng giải quyết vấn đề tải giá trị thông qua một phương pháp Hail Mary, chúng ta sẽ thất bại. Cần phải có ai đó mạnh tay hơn.

26. Christiano (2012).

27. AI mà chúng ta xây dựng cũng không cần phải tìm thấy mô hình. Giống như chúng ta, nó có thể lý giải về những gì một định nghĩa ngầm phức tạp như vậy sẽ đòi hỏi (có lẽ bằng cách nhìn vào môi trường của nó và tuân theo cùng một kiểu lý luận mà chúng ta sẽ áp dụng).

28. So sánh với Chương 9 và 11.

29. Chẳng hạn, MDMA có thể tạm thời tăng sự đồng cảm; oxytocin có thể tạm thời làm tăng lòng tin (Vollenweider và cộng sự 1998; Bartz và cộng sự 2011). Tuy nhiên, các hiệu ứng có vẻ khá đa dạng và phụ thuộc bối cảnh.

30. Các tác tử được tăng cường có thể bị tiêu diệt hoặc tạm dừng, được đặt lại về trạng thái trước đó hoặc bị lấy đi năng lực và không được nhận bất kỳ cải tiến nào nữa, cho đến khi toàn bộ hệ thống đạt đến trạng thái an toàn và trưởng thành hơn, trong đó các yếu tố “lừa đảo” trước đây không còn là mối đe dọa trên toàn hệ thống.

31. Vấn đề cũng có thể ít rõ ràng hơn trong một xã hội tương lai của người sinh học, một người có khả năng tiếp cận kỹ thuật giám sát tiên tiến hoặc kỹ thuật y sinh để điều khiển tâm lý, hay đủ giàu để có được vô số chuyên gia bảo mật để điều tra thường xuyên công dân (và điều tra lẫn nhau).

32. So sánh với Armstrong (2007) và Shulman (2010b).

33. Một câu hỏi mở là ở mức độ nào giám sát viên cấp độ n cần theo dõi không chỉ giám sát viên cấp độ $(n - 1)$ của họ, mà cả giám sát viên cấp độ $(n - 2)$, để biết rằng các tác tử cấp độ $(n - 1)$ đang làm việc đúng cách. Và để biết rằng các tác tử cấp $(n - 1)$ đã quản lý thành công các tác tử cấp $(n - 2)$, tác tử cấp n liệu có cần giám sát các tác tử cấp $(n - 3)$ không?

34. Cách tiếp cận này đứng giữa ranh giới giữa lựa chọn động lực và kiểm soát năng lực. Về mặt kỹ thuật, phần của bố cục mà trong đó, con người điều khiển một giám sát viên phần mềm được coi là kiểm soát năng lực, trong khi phần bố cục có các lớp tác tử

phần mềm kiểm soát lẫn nhau là lựa chọn động lực (khi mà chính bố cục là thứ quyết định xu hướng của hệ thống).

35. Trên thực tế, nhiều chi phí khác đáng được xem xét nhưng không thể đưa ra ở đây. Ví dụ, bất kỳ tác tử nào bị buộc tội cai trị một hệ thống phân cấp như vậy có thể bị hỏng hoặc bị gỡ bỏ bởi quyền lực.

36. Để sự đảm bảo này có hiệu quả, nó phải được thực hiện với thiện chí. Điều này sẽ loại trừ một số loại thao tác của năng lực cảm xúc và ra quyết định của bản giả lập mà trong trường hợp khác, có thể được sử dụng (ví dụ) để “cài đặt” một nỗi sợ bị tạm dừng hoặc ngăn bản giả lập xem xét các lựa chọn của mình một cách hợp lý.

37. Ví dụ, xem Brinton (1965); Goldstone (1980, 2001). (Tiến bộ khoa học xã hội đối với những câu hỏi này có thể tạo ra một món quà tuyệt vời cho những người tuyệt vọng trên thế giới, những người có thể sử dụng các mô hình dự đoán chính xác hơn về tình trạng bất ổn xã hội để tối ưu hóa các chiến lược kiểm soát dân số của họ và nhẹ nhàng dập tắt các cuộc nổi dậy từ trứng nước.)

38. So sánh với Bostrom (2011a, 2009b).

39. Với một hệ thống hoàn toàn nhân tạo, có thể có được một số lợi thế của cấu trúc thể chế mà không thực sự tạo ra các tác tử con độc lập. Một hệ thống có thể kết hợp nhiều quan điểm vào quá trình ra quyết định của nó mà không “vũ trang” đầy đủ cho các quan điểm đó với các năng lực nhận thức cần thiết cho cơ quan độc lập. Tuy nhiên, có thể rất khó để thực hiện đầy đủ tính năng “quan sát về hậu quả hành vi của một thay đổi được đề xuất, và hoàn nguyên phiên bản trước nếu hậu quả tỏ ra không mong muốn từ góc nhìn dự kiến”

được mô tả trong sách trong một hệ thống không bao gồm các tác tử con.

CHƯƠNG 13: LỰA CHỌN TIÊU CHÍ CHỌN LỰA

1. Một nghiên cứu gần đây tiến hành trên triết gia đã cho thấy tỷ lệ người được hỏi “chấp nhận hoặc thiên về” các quan điểm khác nhau. Về đạo đức học chuẩn tắc, kết quả là đạo lý học 25,9%; hệ quả luận 23,6%; đức hạnh luận 18,2%. Về siêu đạo đức (metaethics), kết quả là chủ nghĩa hiện thực đạo đức 56,4%; chủ nghĩa hiện thực chống đạo đức 27,7%. Về đánh giá đạo đức: thuyết nhận thức 65,7%; thuyết phi nhận thức 17,0% (Bourget và Chalmer 2009).

2. Pinker (2011).

3. Xem thảo luận trong Shulman và cộng sự (2009).

4. Moore (2011).

5. Bostrom (2006b).

6. Bostrom (2009b).

7. Bostrom (2011a).

8. Chính xác hơn, chúng ta nên làm theo quan điểm của nó, ngoại trừ về những chủ đề mà chúng ta có lý do chính đáng để cho rằng niềm tin của chúng ta chính xác hơn. Ví dụ, chúng ta có thể biết nhiều hơn về những gì mình đang nghĩ tại một thời điểm cụ thể so với siêu trí tuệ, nếu nó không thể quét não của chúng ta. Tuy nhiên, chúng ta có thể bỏ qua phẩm chất này nếu cho rằng siêu trí tuệ có quyền truy cập vào các ý kiến của mình; sau đó chúng ta cũng có thể giao cho siêu trí tuệ nhiệm vụ đánh giá xem khi nào những ý kiến của chúng ta là đáng tin cậy. (Có thể vẫn còn một số trường hợp đặc biệt, liên quan đến thông tin mang tính chỉ biệt, cần được xử lý riêng biệt, chẳng hạn như bằng việc yêu cầu siêu trí tuệ giải thích cho chúng ta

điều gì đáng tin từ quan điểm của chúng ta.) Xem bài viết về sự chứng nhận và chuyên gia nhận thức (testimony and epistemic authority) trong các tài liệu triết học đang phát triển trong Elga (2007).

9. Yudkowsky (2004). Xem thêm Mijic (2010).

10. Ví dụ, David Lewis đã đề xuất một lý thuyết giá trị khuynh hướng (dispositional theory of value), rằng một X nào đó là giá trị cho A nếu và chỉ nếu A muốn X, nếu A hoàn toàn có lý do và quen thuộc với X một cách lý tưởng (Smith và cộng sự 1989). Những ý tưởng tương tự đã được đưa ra trước đó; ví dụ, xem Sen và Williams (1982), Railton (1986) và Sidgwick và Jones (2010). Tương tự, một quan điểm triết học phổ biến, phương pháp quân bình từ suy tưởng, đề xuất một quá trình điều chỉnh lẫn nhau lặp đi lặp lại giữa: trực giác của chúng ta về các trường hợp cụ thể, các quy tắc chung mà chúng ta nghĩ là chi phối các trường hợp này, và các nguyên tắc mà theo đó chúng ta nghĩ những yếu tố này cần được sửa đổi; để đạt được một hệ thống chặt chẽ hơn; Ví dụ, xem Rawls (1971) và Goodman (1954).

11. Có lẽ chủ ý ở đây là khi AI hành động để ngăn chặn những thảm họa như vậy, nó nên làm điều đó với một “cú chạm” nhẹ nhất có thể, nghĩa là ngăn chặn thảm họa nhưng không ảnh hưởng quá nhiều đến cách mọi thứ diễn ra đối với nhân loại ở những khía cạnh khác.

12. Yudkowsky (2004).

13. Rebecca Roache, trao đổi cá nhân.

14. Ba nguyên tắc là: “Bảo vệ con người, tương lai nhân loại và bản chất con người lý tưởng” (con người lý tưởng – humane là thứ mà chúng ta mong muốn trở thành, khác với con người – human,

chính là chúng ta hiện tại); “Nhân loại không nên dành phần còn lại của sự vĩnh hằng để ước muốn một cách tuyệt vọng rằng các lập trình viên đã làm một điều gì đó khác biệt”; và “Giúp đỡ mọi người”.

15. Một số nhóm tôn giáo chú trọng mạnh mẽ vào niềm tin, thứ đối lập với lý trí, thứ mà họ có thể coi là không đủ để đạt được những hiểu biết tinh thần thiết yếu (ngay cả khi thứ đó ở dạng lý tưởng nhất trong giả thuyết và ngay cả sau khi đã hăng hái và cởi mở học hết từng dòng trong Kinh thánh, Khải huyền và Giải kinh). Những người có quan điểm như vậy có thể không coi CEV là một hướng dẫn tối ưu cho việc ra quyết định (mặc dù họ vẫn có thể thích nó hơn các hướng dẫn không hoàn hảo khác có thể làm theo nếu phương pháp CEV được thực hiện).

16. Một AI hoạt động như một lực lượng tự nhiên tiềm ẩn để điều chỉnh các tương tác của con người đã được gọi là một “Sysop”, một loại “hệ điều hành” cho phần vật chất bị chiếm giữ bởi nền văn minh của loài người. Xem Yudkowsky (2001).

17. Là “Có thể”, vì với điều kiện ý chí ngoại suy kết hợp của con người mong muốn không mở rộng việc cân nhắc đạo đức cho các thực thể này, có lẽ việc liệu những thực thể đó có thực sự có vị thế đạo đức hay không (mặc dù hiện tại chúng có vẻ rất hợp lý) là khá đáng nghi ngờ. “Có khả năng”, vì ngay cả khi một cuộc “bỏ phiếu” chống ngăn chặn động lực CEV trực tiếp bảo vệ những người bên ngoài này, vẫn có một khả năng rằng trong bất kỳ quy tắc nền tảng nào còn sót lại sau khi động lực ban đầu được thực hiện, những cá nhân có mong muốn được tôn trọng và muốn lợi ích của người bên ngoài được bảo vệ có thể thương lượng thành công để đạt được kết quả này (nhưng phải từ bỏ một lượng tài nguyên của chính mình).

Liệu điều này có khả thi hay không có thể một phần phụ thuộc vào việc liệu kết quả của động lực CEV có phải là một bộ quy tắc cơ bản giúp đạt được các giải pháp được đàm phán cho các vấn đề thuộc loại này hay không (có thể đòi hỏi các điều khoản để khắc phục các vấn đề thương lượng chiến lược).

18. Các cá nhân đóng góp tích cực để hiện thực hóa một siêu trí tuệ an toàn và có lợi có thể xứng đáng với một phần thưởng đặc biệt cho công sức của họ, mặc dù có một sứ mệnh gần như độc quyền để xác định khuynh hướng của tài nguyên vũ trụ. Tuy nhiên, khái niệm về việc tất cả mọi người nhận được một phần bằng nhau trong cơ sở ngoại suy của chúng ta là một tiêu điểm (Schelling point, hoặc focal point) mỹ mãn đến mức không nên bỏ qua một cách dễ dàng. Trong mọi trường hợp, có một cách gián tiếp mà với đó, đức hạnh có thể được khen thưởng: cụ thể, chính CEV có thể chỉ ra rằng những người tốt đã gắng sức thay mặt nhân loại nên được công nhận một cách phù hợp. Điều này có thể xảy ra mà không cần những người như vậy phải có được bất kỳ trọng số đặc biệt nào trong cơ sở ngoại suy nếu CEV của chúng ta chấp thuận (theo nghĩa là đưa ra ít nhất một trọng số khác 0) một nguyên tắc về sự xứng đáng.

19. Bostrom và cộng sự (2013).

20. Ở chừng mực mà một ý nghĩa chung (đủ xác định) đang được thể hiện khi chúng ta đưa ra các đánh giá đạo đức, một siêu trí tuệ có thể hiểu được ý nghĩa đó là gì. Và trong phạm vi mà các đánh giá đạo đức có “giá trị chân lý” (truth-apt, đặc tính của mệnh đề cho phép nó là đúng hoặc sai), siêu trí tuệ có thể tìm ra đánh giá nào của dạng mệnh đề “Bây giờ tác tử X phải làm việc Φ ” là đúng. Ít nhất, nó nên giỏi hơn chúng ta trong nhiệm vụ này.

Một AI ban đầu thiếu khả năng nhận thức đạo đức như vậy sẽ có khả năng đó sau này nếu nó có siêu năng lực khuếch đại trí tuệ. Một cách mà AI có thể làm là đảo ngược kỹ nghệ những suy nghĩ đạo đức của não người, và sau đó thực hiện một quy trình tương tự nhưng chạy nhanh hơn, cung cấp cho nó thông tin thực tế chính xác hơn, và nhiều lợi ích khác.

21. Vì chúng ta không chắc chắn về siêu đạo đức, nên có một câu hỏi về việc AI sẽ làm gì nếu các điều kiện tiên quyết cho MR không đạt được. Một lựa chọn là quy định rằng AI tự tắt nếu nó gán xác suất đủ cao cho việc nhận thức luận đạo đức là sai, hoặc cho việc không có sự thật đạo đức không liên (non-relative) quan phù hợp. Ngoài ra, chúng ta có thể khiến AI hoàn nguyên về một số phương pháp thay thế, chẳng hạn như CEV.

Chúng ta có thể tinh chỉnh đề xuất MR rõ ràng hơn về những gì sẽ được thực hiện trong các trường hợp mơ hồ hoặc suy biến khác nhau. Chẳng hạn, nếu lý thuyết lỗi là đúng (và do đó, tất cả các đánh giá đạo đức tích cực của dạng mệnh đề “Bây giờ tôi phải làm việc Φ ” là sai), thì chiến lược dự phòng (như tự tắt) sẽ được đưa ra. Chúng ta cũng có thể chỉ định những gì sẽ xảy ra nếu có nhiều hành động khả thi, mỗi hành động đều đúng về đạo đức. Ví dụ, chúng ta có thể nói rằng trong những trường hợp như vậy, AI sẽ thực hiện (một trong) những hành động được phép mà phép ngoại suy tập thể của loài người sẽ ưu tiên. Chúng ta cũng có thể quy định điều gì sẽ xảy ra nếu lý thuyết đạo đức thực sự không sử dụng các cụm như “đúng đắn về đạo đức” trong từ vựng cơ bản của nó. Chẳng hạn, một lý thuyết hệ quả luận có thể cho rằng một số hành động tốt hơn những hành động khác, nhưng không có ngưỡng cụ thể nào tương ứng với

khái niệm hành động là đúng về đạo đức. Sau đó chúng ta có thể nói rằng nếu một lý thuyết như vậy là chính xác, MR nên thực hiện một trong những hành động khả thi tốt nhất về đạo đức, nếu có; hoặc, nếu có vô số hành động khả thi, sao cho bất kỳ hành động nào trong số đó đều có một hành động tốt hơn, thì có lẽ MR có thể chọn bất kỳ hành động nào ít nhất là tốt hơn đáng kể so với hành động tốt nhất mà bất kỳ con người nào đã chọn trong tình huống tương tự, nếu một hành động như vậy là khả thi (hoặc nếu không thì là một hành động ít nhất cũng tốt như hành động tốt nhất mà con người sẽ thực hiện.)

Một vài điểm chung nên được lưu ý khi nghĩ về cách đề xuất MR có thể được tinh chỉnh. Đầu tiên, chúng ta có thể bắt đầu một cách bảo thủ, sử dụng tùy chọn dự phòng để bao quát gần như tất cả các trường hợp bất ngờ và chỉ sử dụng tùy chọn “đúng đắn về đạo đức” trong những trường hợp mà chúng ta cảm thấy hoàn toàn hiểu được. Thứ hai, chúng ta có thể thêm bộ điều biến chung vào đề xuất MR rằng nó sẽ được “diễn dịch một cách hào phóng và sửa đổi như chúng ta sẽ làm nếu đã suy nghĩ kỹ hơn về nó trước khi viết.”

22. Trong số các thuật ngữ này, “kiến thức” dường như là thứ dễ phân tích nhất (về mặt lý thuyết thông tin). Tuy nhiên, để thể hiện những gì con người biết về một thứ gì đó, AI có thể cần một tập hợp các phép biểu diễn tinh vi liên quan đến các thuộc tính tâm lý phức tạp. Một con người không “biết” được hết toàn bộ thông tin được lưu trữ trong não của mình.

23. Một chỉ số cho thấy các thuật ngữ trong CEV rõ ràng hơn (một chút) là: sẽ tính là tiến bộ về triết học nếu chúng ta có thể phân tích sự đúng đắn về đạo đức bằng các thuật ngữ giống như trong CEV.

Trên thực tế, một trong những mạch chính trong siêu đạo đức – lý thuyết người quan sát lý tưởng – có thể làm được điều đó. Ví dụ, xem Smith và cộng sự (1989).

24. Điều này đòi hỏi phải đối mặt với vấn đề của sự không chắc chắn quy chuẩn cơ bản. Có thể chỉ ra rằng không phải lúc nào cũng thích hợp để hành động theo lý thuyết đạo đức có xác suất đúng cao nhất, và cũng có thể chỉ ra rằng không phải lúc nào cũng thích hợp để thực hiện hành động có xác suất đúng cao nhất. Một số cách đánh đổi xác suất với “mức độ sai lầm” hoặc mức độ nghiêm trọng của các vấn đề đang bị đe dọa dường như là cần thiết. Đối với một số ý tưởng theo hướng này, xem Bostrom (2009a).

25. Thậm chí có thể lập luận rằng đó là một điều kiện phù hợp cho bất kỳ sự lý giải nào về khái niệm quyền đạo đức mà nó giải thích cho việc Joe Sixpack có thể biết đôi chút về đúng và sai.

26. Điều đúng đắn về đạo đức mà chúng ta phải làm là xây dựng một AI thực hiện MR là không rõ ràng, ngay cả khi chúng ta cho rằng chính AI sẽ luôn hành động một cách đạo đức. Có lẽ sẽ là một sự ngạo mạn đáng lên án khi chúng ta xây dựng một AI như vậy (đặc biệt là vì nhiều người có thể không chấp nhận dự án đó). Vấn đề này có thể được giải quyết một phần bằng cách điều chỉnh đề xuất MR. Giả sử, chúng ta quy định rằng AI nên hành động (làm những gì đúng đắn về đạo đức) chỉ khi việc người ta tạo ra AI này ngay từ đầu là đúng đắn về đạo đức; nếu không thì nó nên tự tắt. Thật khó để thấy làm sao chúng ta phạm phải bất kỳ sai lầm nghiêm trọng nào về đạo đức khi tạo ra loại AI đó, vì nếu việc chúng ta tạo ra nó là sai, thì hậu quả duy nhất là AI được tạo ra sẽ tự tắt ngay lập tức nếu AI đã không phạm tội ác tâm trí cho đến thời điểm đó. (Tuy nhiên, chúng ta

có thể đã hành động sai lầm, chẳng hạn, vì đã không nắm bắt được cơ hội để xây dựng một AI nào đó khác.)

Một vấn đề thứ hai là những hành động vượt quá sự đòi hỏi. Giả sử có rất nhiều hành động mà AI có thể thực hiện, mỗi hành động đó sẽ đúng về đạo đức (tức là được phép về đạo đức), nhưng một số trong số đó tốt hơn về đạo đức. Một lựa chọn là khiến AI đặt mục tiêu chọn hành động tốt nhất về đạo đức trong bất kỳ tình huống nào như vậy (hoặc một trong những hành động tốt nhất, trong trường hợp có nhiều hành động tốt như nhau). Một lựa chọn khác là khiến AI lựa chọn trong số các hành động được phép về đạo đức một hành động thỏa mãn tối đa một số ước nguyện (phi đạo đức) khác. Ví dụ: AI có thể chọn, trong số các hành động được phép về đạo đức, hành động mà CEV của chúng ta muốn thực hiện. Một AI như vậy, trong khi không bao giờ làm bất cứ điều gì không được phép về đạo đức, có thể bảo vệ lợi ích của chúng ta nhiều hơn một AI làm những gì tốt nhất về đạo đức.

27. Khi AI đánh giá mức độ được phép về đạo đức trong hành động tạo ra AI của chúng ta, nó sẽ diễn giải mức độ được phép theo nghĩa khách quan của nó. Theo một nghĩa thông thường về mức độ được phép về đạo đức, một bác sĩ hành động cho phép về đạo đức khi kê đơn một loại thuốc mà mình tin rằng sẽ chữa khỏi cho bệnh nhân ngay cả khi bệnh nhân, mà bác sĩ không hề quen biết, bị dị ứng với thuốc và tử vong. Tập trung vào mức độ được phép về đạo đức khách quan tận dụng vị thế nhận thức vượt trội có lẽ là vượt trội của AI.

28. Trực tiếp hơn, nó phụ thuộc vào niềm tin của AI về việc lý thuyết đạo đức nào là đúng (hay chính xác hơn là phân phối xác suất

của nó đối với các lý thuyết đạo đức).

29. Thật khó tưởng tượng những sinh vật khả thể về mặt vật lý này có thể tuyệt vời đến đâu. Xem Bostrom (2008c) cho một nỗ lực “thi vị” để truyền đạt ý nghĩa về điều này. Xem Bostrom (2008b) về lập luận rằng một số khả năng như vậy có thể tốt cho chúng ta, tốt cho con người hiện tại.

30. Xem ra việc tiến cử một đề xuất nếu ai đó nghĩ rằng một đề xuất nào đó khác là tốt hơn sẽ là lừa đảo hoặc thao túng. Nhưng người ta có thể thúc đẩy nó mà vẫn tránh được sự không trung thực. Ví dụ, người ta có thể tự do thừa nhận tính ưu việt của đề xuất lý tưởng trong khi vẫn tiến cử đề xuất phi lý tưởng như là sự thỏa hiệp dễ đạt được nhất.

31. Hoặc một số cụm từ đánh giá tích cực khác, chẳng hạn như “tốt”, “tuyệt” hay “mỹ mãn”.

32. Điều này lặp lại một nguyên tắc trong thiết kế phần mềm được biết đến với tên là “Hãy làm điều tôi muốn”, viết tắt là DWIM (Do What I Mean). Xem Teitelman (1966).

33. Nội dung mục tiêu, lý thuyết quyết định và nhận thức luận là ba khía cạnh cần được làm sáng tỏ; nhưng chúng tôi không có ý định đặt ra câu hỏi liệu có cần phân tách thành ba thành phần riêng biệt này hay không.

34. Một dự án đạo đức có lẽ nên phân bổ ít nhất là một phần khiêm tốn của lợi ích cuối cùng mà siêu trí tuệ tạo ra như là phần thưởng đặc biệt cho những người đã cố gắng theo những cách được phép về đạo đức vào thành công của dự án. Phân bổ một phần lớn cho kế hoạch gói phần thưởng sẽ là không đáng kể. Nó tương tự như một tổ chức từ thiện dành 90% thu nhập của mình cho các

khoản thưởng hiệu suất cho những người gây quỹ và cho các chiến dịch quảng cáo để tăng quyên góp.

35. Làm thế nào người chết có thể được khen thưởng? Người ta có thể nghĩ về một số khả năng. Ở cấp thấp, có thể có các hoạt động tưởng niệm và tượng đài, đó sẽ là một phần thưởng cho người mong muốn danh tiếng sau khi chết. Người chết cũng có thể có những ưu tiên khác về tương lai có thể được vinh danh, chẳng hạn như liên quan đến văn hóa, nghệ thuật, xây dựng hoặc môi trường tự nhiên. Hơn nữa, hầu hết mọi người quan tâm đến con cháu của họ, và các đặc quyền đặc biệt có thể được trao cho con cháu của những người đóng góp. Cụ thể hơn, siêu trí tuệ có thể tạo ra các mô phỏng tương đối trung thực của một số người quá cố – các bản mô phỏng có ý thức và sẽ giống với bản gốc đủ để được coi là một hình thức sinh tồn (theo ít nhất là tiêu chí của một số người). Điều này có lẽ sẽ dễ dàng hơn cho những người được bảo quản trong dung dịch huyền phù lạnh; nhưng có lẽ đối với một siêu trí tuệ, sẽ không thể tái tạo một cái gì đó tương tự người gốc từ các hồ sơ được bảo quản khác như thư từ, ấn phẩm, tài liệu nghe nhìn và hồ sơ kỹ thuật số, hoặc ký ức cá nhân của những người còn sống khác. Một siêu trí tuệ cũng có thể nghĩ về một số khả năng không dễ dàng xảy ra với chúng ta.

36. Về thuyết cướp đoạt theo kiểu Pascal (Pascalian mugging), xem Bostrom (2009b). Để phân tích các vấn đề liên quan đến các mức độ thỏa dụng vô hạn, xem Bostrom (2011a). Về tính không chắc chắn quy phạm cơ bản, Ví dụ, xem Bostrom (2009a).

37. Ví dụ: Price (1991); Joyce (1999); Luxcher (2006); Yudkowsky (2010); Dai (2009).

38. Ví dụ: Bostrom (2009a).

39. Cũng có thể hình dung rằng việc sử dụng phương pháp quy chuẩn gián tiếp để chỉ định nội dung mục tiêu của AI sẽ giảm thiểu các vấn đề có thể phát sinh từ một lý thuyết quyết định được chỉ định không chính xác. Hãy xem xét, ví dụ, phương pháp CEV. Nếu được triển khai tốt, nó có khả năng bù lại ít nhất một số lỗi trong đặc điểm kỹ thuật của lý thuyết quyết định của AI. Việc triển khai có thể cho phép các giá trị mà ý chí ngoại suy kết hợp của chúng ta muốn AI theo đuổi phụ thuộc vào lý thuyết quyết định của AI. Nếu bản thể được lý tưởng hóa của chúng ta biết rằng họ đang tạo ra các bản đặc tả có giá trị cho một AI đang sử dụng một loại lý thuyết quyết định cụ thể, họ có thể điều chỉnh các bản đặc tả giá trị giá trị của mình nhằm làm cho AI hành xử một cách lành tính mặc dù lý thuyết quyết định của nó bị biến dạng (giống như người ta có thể hủy bỏ hiệu ứng làm biến dạng của một ống kính bằng cách đặt một ống kính khác trước ống kính đó để làm biến dạng ngược lại).

40. Một số hệ thống nhận thức luận có thể, một cách toàn diện, không có nền tảng riêng biệt. Trong trường hợp đó, sự kế thừa thể chế không phải là một bộ nguyên tắc riêng biệt, mà đúng hơn, là một điểm khởi đầu về nhận thức thể hiện một số khuynh hướng nhất định để đáp ứng với những bằng chứng sắp tới.

41. Ví dụ, xem vấn đề biến dạng được thảo luận trong Bostrom (2011a).

42. Chẳng hạn, một vấn đề gây tranh cãi trong lý luận nhân học là liệu cái gọi là giả định tự chỉ định (self-indication assumption) có nên được chấp nhận hay không. Giả định tự chỉ định, đại khái, nói rằng từ việc bạn tồn tại, bạn nên suy ra rằng các giả thuyết mà theo đó số

lượng N lớn hơn của những người quan sát cũng tồn tại sẽ được gia tăng xác suất tỷ lệ thuận với N . Xem phần lập luận phản bác nguyên lý này trong thí nghiệm tưởng tượng “Presumptuous Philosopher” (Vị triết gia kiêu ngạo) trong Bostrom (2002a). Xem lập luận ủng hộ nguyên lý này trong Olum (2002); xem phần phản bác với ý kiến ủng hộ đó trong Bostrom và Ćirković (2003). Niềm tin về giả định tự chỉ định có thể ảnh hưởng đến các giả thuyết thực nghiệm khác nhau về mức độ phù hợp chiến lược quan trọng, ví dụ, các cân nhắc như lập luận ngày tận thế Carter-Leslie, lập luận mô phỏng và lập luận “bộ lọc lớn”. Xem Bostrom (2002a, 2003a, Bostrom 2008a); Carter (1983); Ćirković và cộng sự (2010); Hanson (1998d); Leslie (1996); Tegmark và Bostrom (2005). Một điểm tương tự có thể được đưa ra liên quan đến các vấn đề nhức nhối khác trong lý thuyết công suất tính toán, chẳng hạn như liệu sự lựa chọn của lớp tham chiếu có thể được coi là có liên quan đến cụm “người quan sát-khoảnh khắc” hay không, và nếu có thì liên quan ra sao.

43. Ví dụ, xem Howson và Urbach (1993). Ngoài ra còn có một số kết quả thú vị thu hẹp phạm vi tình huống mà trong đó, hai tác tử Bayes có thể bất đồng về lý trí khi ý kiến của chúng là kiến thức phổ biến; xem Aumann (1976) và Hanson (2006).

44. So sánh với khái niệm về “người phân xử cuối cùng” (last judge) của Yudkowsky (2004).

45. Có nhiều vấn đề quan trọng nổi bật trong nhận thức luận và một số đã được đề cập trong sách. Vấn đề ở đây là chúng ta có thể không cần phải thực hiện một cách hoàn hảo mọi giải pháp để đạt được một kết quả mà thực tế không khác kết quả tốt nhất là bao. Một

mô hình hỗn hợp (kết hợp nhiều tiên nghiệm khác nhau) có thể có hiệu quả.

CHƯƠNG 14: BỨC TRANH CHIẾN LƯỢC

1. Nguyên tắc này được giới thiệu trong Bostrom (2009b, 190), trong đó cũng lưu ý rằng nó không phải là một lối diễn đạt lặp từ thừa thãi. Để trực quan, hãy hình dung một chiếc hộp với khối lượng lớn nhưng hữu hạn, đại diện cho không gian của các khả năng cơ bản có thể có được thông qua một số công nghệ khả thể. Hãy tưởng tượng cát được đổ vào hộp này, đại diện cho nỗ lực nghiên cứu. Cách bạn đổ cát sẽ quyết định vị trí mà ở đó cát chất thành đống. Nhưng nếu bạn tiếp tục đổ, toàn bộ không gian trong hộp cuối cùng sẽ được lấp đầy.

2. Bostrom (2002b).

3. Đây không phải là quan điểm thường gặp về chính sách khoa học. Harvey Averch mô tả chính sách khoa học và công nghệ ở Mỹ trong khoảng thời gian từ năm 1945 đến 1984 là tập trung vào các cuộc tranh luận về mức đầu tư công tối ưu vào khoa học và công nghệ, và vào mức độ mà chính phủ nên cố gắng “chọn người thắng cuộc” để đạt được mức tăng lớn nhất trong sự thịnh vượng kinh tế và sức mạnh quân sự của quốc gia. Trong các tính toán này, tiến bộ công nghệ luôn được coi là tốt. Nhưng Averch cũng mô tả sự trỗi dậy của các quan điểm trọng yếu, vốn luôn nghi ngờ giả thuyết “tiến bộ luôn là tốt” (Averch 1985). Xem thêm Graham (1997).

4. Bostrom (2002b).

5. Điều này là tất nhiên là không hề thừa thãi. Người ta có thể tưởng tượng ra trường hợp mà ở đó có một trật tự phát triển khác. Có thể lập luận việc đối mặt trước với một thách thức ít khó khăn

hơn sẽ tốt hơn cho nhân loại, nghĩa là sự phát triển của công nghệ nano, với lý do điều này sẽ buộc chúng ta phải phát triển các thể chế tốt hơn, phối hợp quốc tế tích cực hơn và trưởng thành hơn trong suy nghĩ về chiến lược toàn cầu. Có lẽ chúng ta sẽ có nhiều khả năng đương đầu với một thách thức với mối đe dọa dễ hiểu hơn về mặt siêu hình so với siêu trí tuệ máy. Công nghệ nano (hoặc sinh học tổng hợp, hoặc bất kỳ thử thách nào mà chúng ta phải đối mặt trước tiên) có thể đóng vai trò là bước đệm giúp chúng ta đạt đến mức năng lực cần thiết để đối phó với thách thức siêu trí tuệ cấp cao hơn. Một lập luận như vậy sẽ phải được đánh giá trên cơ sở từng trường hợp. Ví dụ, trong trường hợp của công nghệ nano, người ta sẽ phải xem xét các điều kiện khả thi khác nhau, chẳng hạn như tăng hiệu suất phần cứng từ các chất nền tính toán được chế tạo bằng sợi nano; các tác động của một nguồn vốn vật lý thấp phục vụ sản xuất đối với sự tăng trưởng kinh tế; sự phát triển của công nghệ giám sát tinh vi; khả năng một thể đơn nhất có thể xuất hiện bất kể có các tác động trực tiếp hay gián tiếp của một đột phá công nghệ nano; và tính khả thi cao hơn của các phương pháp tiếp cận hình thái thần kinh và giả lập hoàn chỉnh não bộ đối với trí tuệ của máy. Việc xem xét tất cả các vấn đề này (hoặc các vấn đề song song có thể phát sinh đối với các công nghệ gây rủi ro diệt vong khác) là vượt quá phạm vi nghiên cứu của chúng ta. Ở đây chúng tôi chỉ đưa ra trường hợp hiển nhiên nhằm ủng hộ thứ tự phát triển mà siêu trí tuệ đi đầu, trong khi nhấn mạnh rằng có những sự phức tạp có thể làm thay đổi đánh giá sơ bộ này trong một số trường hợp.

6. Pinker (2011); Wright (2001).

7. Chúng ta có thể rất muốn cho rằng giả thuyết về việc mọi thứ đã tăng tốc để trở nên vô nghĩa vì (thoạt nhìn) dường như không có bất kỳ hậu quả quan sát nào; nhưng hãy xem, ví dụ, Shoemaker (1969).

8. Mức độ sẵn sàng không được đo lường bằng số lượng nỗ lực dành cho các hoạt động chuẩn bị, mà bằng việc các điều kiện thực sự được định hình thuận lợi như thế nào, và các nhà ra quyết định chủ chốt sẵn sàng hành động ra sao.

9. Mức độ tin cậy quốc tế trong giai đoạn dẫn đến sự bùng nổ trí tuệ cũng có thể là một nhân tố. Chúng ta sẽ xem xét điều này trong phần “Hợp tác”.

10. Theo giai thoại, có vẻ như những người hiện đang quan tâm nghiêm túc đến vấn đề kiểm soát được lấy mẫu một cách không tương xứng từ một thái cực của bố cục phân phối thông tin, mặc dù có thể có những giải thích khác. Nếu lĩnh vực này trở thành “mốt”, chắc chắn nó sẽ có đầy rẫy những kẻ tầm thường và lập dị.

11. Tôi mượn Carl Shulman thuật ngữ này.

12. Trí tuệ máy phải giống não bộ đến đâu để được tính là một bản giả lập hoàn chỉnh não bộ chứ không phải là một AI hình thái thần kinh? Yếu tố quyết định có liên quan có thể là liệu hệ thống có tái tạo các giá trị hoặc toàn bộ xu hướng nhận thức và đánh giá của một cá nhân cụ thể hoặc một con người chung chung hay không, bởi vì điều này sẽ tạo ra sự khác biệt cho vấn đề kiểm soát. Bản giả lập cần có độ trung thực khá cao để sao chép được các tính chất này.

13. Độ lớn của sự tăng cường tất nhiên sẽ phụ thuộc vào độ lớn và nguồn gốc phát sinh sự thúc đẩy. Có thể không có sự tăng cường nào cho khoa học thần kinh nếu tất cả các nguồn lực đầu tư vào

ngiên cứu giả lập hoàn chỉnh não bộ được rút ra từ nghiên cứu khoa học thần kinh thông thường, trừ khi sự tập trung vào nghiên cứu giả lập tình cờ là một cách hiệu quả hơn để tăng cường khoa học thần kinh so với danh mục đầu tư mặc định vào nghiên cứu khoa học thần kinh.

14. Xem Drexler (1986, 242). Drexler (trao đổi cá nhân) xác nhận rằng sự tái cấu trúc này tương ứng với lập luận mà ông đang tìm cách trình bày. Rõ ràng, một số tiền đề ngầm sẽ phải được thêm vào nếu người ta muốn đưa ra lập luận dưới dạng một chuỗi lý luận có giá trị về mặt suy diễn. (Lưu ý rằng Drexler không xác nhận “loại luận điểm có liên quan” được thảo luận bên dưới.)

15. Có lẽ chúng ta không nên đón nhận những thảm họa nhỏ phòng khi chúng làm tăng sự cảnh giác của chúng ta đến mức khiến chúng ta ngăn chặn những thảm họa quy mô trung bình, vốn cần thiết để khiến chúng ta thực hiện các biện pháp phòng ngừa mạnh mẽ nhằm ngăn chặn thảm họa diệt vong? (Và tất nhiên, giống như với các hệ thống miễn dịch sinh học, chúng ta cũng cần quan tâm đến các phản ứng quá mức, tương tự như dị ứng và rối loạn tự miễn dịch.)

16. So sánh với Lenman (2000); Burch-Brown (2014).

17. So sánh với Bostrom (2007).

18. Lưu ý rằng luận điểm này tập trung vào thứ tự thay vì thời gian của các sự kiện có liên quan. Làm cho siêu trí tuệ xảy ra sớm hơn chỉ giúp tránh các rủi ro diệt vong khác nếu sự can thiệp làm thay đổi trình tự của các phát triển chính: ví dụ, bằng cách làm cho siêu trí tuệ xảy ra trước khi đạt được các mốc quan trọng khác nhau trong công nghệ nano hoặc sinh học tổng hợp.

19. Nếu việc giải quyết vấn đề kiểm soát là cực kỳ khó khăn khi so với việc giải quyết vấn đề hiệu năng trong trí tuệ của máy và nếu khả năng của dự án chỉ có mức độ tương quan yếu với quy mô dự án, có thể sẽ tốt hơn khi một dự án nhỏ đến đích trước, cụ thể là nếu sự chênh lệch về khả năng là lớn hơn giữa các dự án nhỏ hơn. Trong tình huống như vậy, ngay cả khi các dự án nhỏ hơn có năng lực trung bình kém hơn các dự án lớn hơn, thì ít có khả năng một dự án nhỏ nhất định sẽ có mức độ năng lực rất cao, cần thiết để giải quyết vấn đề kiểm soát.

20. Không thể phủ nhận rằng người ta có thể tưởng tượng ra các công cụ có thể thúc đẩy thảo luận toàn cầu và sẽ được hưởng lợi, hoặc thậm chí đòi hỏi, sự tiến bộ hơn nữa trong phần cứng, ví dụ như dịch thuật chất lượng cao, công cụ tìm kiếm tốt hơn, phổ cập tiếp cận với điện thoại thông minh, môi trường thực tế ảo hấp dẫn cho giao tiếp xã hội, vv.

21. Đầu tư vào công nghệ giả lập có thể đẩy nhanh tiến độ phát triển theo hướng giả lập hoàn chỉnh não bộ không chỉ trực tiếp (thông qua bất kỳ sản phẩm kỹ thuật nào được sản xuất) mà còn bằng cách tạo ra một nhóm ủng hộ thúc đẩy tài trợ nhiều hơn và tăng khả năng hiển thị và độ tin cậy của giả lập hoàn chỉnh não bộ (WBE).

22. Giá trị kỳ vọng sẽ mất bao nhiêu nếu tương lai được định hình bởi những ham muốn của một con người ngẫu nhiên thay vì (một sự “chồng chập” phù hợp) những ham muốn của toàn nhân loại? Điều này có thể phụ thuộc vào việc chúng ta sử dụng tiêu chuẩn đánh giá nào, và cả việc những mong muốn trên đã được lý tưởng hóa hay chỉ ở mức thô trọng.

23. Ví dụ, trong khi con người giao tiếp chậm qua ngôn ngữ, AI có thể được thiết kế sao cho các phiên bản của cùng một chương trình có thể dễ dàng và nhanh chóng chuyển cả thông tin và kỹ năng cho nhau. Tâm trí máy móc được thiết kế từ ban đầu có thể loại bỏ các hệ thống di sản cồng kềnh giúp tổ tiên chúng ta đối phó với các khía cạnh của môi trường tự nhiên vốn không quan trọng trong không gian ảo. Tâm trí kỹ thuật số cũng có thể được thiết kế để tận dụng khả năng xử lý tuần tự nhanh mà bộ não logic sinh học không có, và để dễ dàng cài đặt các module mới với chức năng tối ưu hóa cao (ví dụ: xử lý biểu tượng, nhận dạng mẫu hình, mô phỏng, khai thác dữ liệu và lập kế hoạch). Trí tuệ nhân tạo cũng có thể có những lợi thế phi kỹ thuật đáng kể, như dễ được cấp bằng sáng chế hơn hoặc ít vướng vào sự phức tạp về đạo đức khi sử dụng các dữ liệu tải lên của con người.

24. Nếu p_1 và p_2 là xác suất thất bại ở mỗi bước, thì tổng xác suất thất bại là $p_1 + (1 - p_1)p_2$ vì người ta có thể thất bại chỉ một lần.

25. Tất nhiên, có thể nhóm đi trước sẽ không có khách hàng tiềm năng lớn như vậy và sẽ không thể tạo thành một thể đơn nhất. Cũng có thể một thể đơn nhất sẽ xuất hiện trước AI ngay cả khi không có sự can thiệp của WBE. Khi đó, lý do để ủng hộ kịch bản WBE đến trước đã không còn.

26. Có cách nào để một người ủng hộ WBE tăng tính đặc thù của sự hỗ trợ của mình nhằm tăng tốc WBE trong khi giảm thiểu hiệu ứng lan tỏa sang việc phát triển AI hay không? Thúc đẩy công nghệ quét có lẽ tốt hơn so với việc thúc đẩy mô hình hóa điện toán thần kinh. (Thúc đẩy phần cứng máy tính không thể tạo ra sự khác biệt theo cách này hay cách khác, khi lợi ích thương mại lớn dù sao cũng

khuyến khích tiến bộ trong lĩnh vực đó.) Thúc đẩy công nghệ quét có thể làm tăng khả năng xảy ra kết quả đa cực bằng cách làm cho việc quét ít có khả năng là “điểm tắc nghẽn”, do đó làm tăng cơ hội của việc quần thể giả lập ban đầu sẽ được sao chép từ nhiều mẫu con người thay vì một số lượng nhỏ các mẫu. Tiến bộ trong công nghệ quét cũng khiến nhiều khả năng điểm tắc nghẽn thay vào đó sẽ là phần cứng máy tính, và sẽ có xu hướng làm chậm quá trình cất cánh.

27. AI hình thái thần kinh cũng có thể thiếu các thuộc tính thúc đẩy an toàn khác của giả lập hoàn chỉnh não bộ, chẳng hạn như có các điểm mạnh và điểm yếu về nhận thức tương tự như con người sinh học (chúng ta sẽ sử dụng kinh nghiệm của mình về con người để hình thành một số kỳ vọng về khả năng của hệ thống ở các giai đoạn phát triển khác nhau).

28. Nếu động cơ thúc đẩy WBE của ai đó là làm cho WBE xảy ra trước AI, họ nên nhớ rằng việc tăng tốc WBE sẽ thay đổi chỉ thứ tự cán đích khi thời điểm mặc định của hai con đường hướng tới trí tuệ máy đã nêu là suýt soát và có một chút lợi thế so với AI. Mặt khác, đầu tư vào WBE sẽ khiến WBE xảy ra sớm hơn (giảm thiểu mái hiên phần cứng và thời gian chuẩn bị) nhưng không ảnh hưởng đến trình tự phát triển; hoặc đầu tư như vậy vào WBE sẽ có rất ít hiệu quả (ngoài việc có thể khiến AI xảy ra thậm chí sớm hơn bằng cách kích thích tiến bộ theo hướng AI hình thái thần kinh).

29. Nhận xét về Hanson (2009).

30. Tất nhiên sẽ có một mức độ to lớn và cấp thiết nào đó của rủi ro diệt vong mà ở đó việc trì hoãn rủi ro sẽ được ưu tiên ngay cả từ quan điểm ảnh hưởng đến cá nhân – liệu có nên cho phép những

người hiện được tận hưởng cuộc sống nhiều hơn một chút trước khi hạ màn hay cung cấp thêm thời gian cho các nỗ lực có thể làm giảm nguy hiểm.

31. Giả sử chúng ta có thể thực hiện một số hành động khiến sự bùng nổ trí tuệ xảy ra sớm hơn một năm. Chẳng hạn, những người hiện đang sống trên Trái đất đang chết dần với tỷ lệ 1% mỗi năm và rủi ro tuyệt chủng của con người từ sự bùng nổ trí tuệ là 20% (chỉ là một con số tùy ý cho mục đích minh họa). Từ đó, việc đẩy nhanh tốc độ xuất hiện của sự bùng nổ trí tuệ mỗi năm có thể tương đương (từ quan điểm ảnh hưởng đến cá nhân) một mức tăng rủi ro từ 20% lên 21%, tức là mức rủi ro tăng 5%. Tuy nhiên, đại đa số người còn sống một năm trước khi bắt đầu sự bùng nổ trí tuệ vào thời điểm đó sẽ quan tâm đến việc trì hoãn nếu họ có thể giảm nguy cơ vụ nổ xuống một 1% (vì hầu hết mọi người sẽ nghĩ rằng nguy cơ qua đời của mình trong năm tới sẽ nhỏ hơn 1%, khi hầu hết các trường hợp tử vong đều xảy ra trong một nhóm dân số nhỏ như những người già yếu bệnh tật). Do đó, người ta có thể có một mô hình trong đó mỗi năm dân số bỏ phiếu để trì hoãn sự bùng nổ trí tuệ thêm một năm nữa, để sự bùng nổ trí tuệ không bao giờ xảy ra, mặc dù mọi người từng sống đều đồng ý rằng sẽ tốt hơn nếu sự bùng nổ trí tuệ xảy ra vào một thời điểm nào đó. Hiển nhiên trong thực tế, sự thất bại trong phối hợp, tính dễ đoán hạn chế hoặc sự ưu tiên cho những thứ khác ngoài sự sống còn của cá nhân có khả năng ngăn chặn sự trì hoãn không hồi kết như vậy.

Nếu một người sử dụng hệ số chiết khấu kinh tế tiêu chuẩn thay vì tiêu chuẩn ảnh hưởng đến con người, thì mức độ lợi ích tiềm năng sẽ giảm đi, vì giá trị của những người hiện tại được hưởng cuộc

sống rất dài sau đó được giảm giá mạnh. Hiệu ứng này đặc biệt mạnh nếu hệ số giảm được áp dụng cho thời gian chủ quan của từng cá nhân thay vì thời gian thiên văn. Nếu lợi ích trong tương lai giảm ở mức $x\%$ mỗi năm và mức rủi ro diệt vong từ các nguồn khác là $y\%$ mỗi năm, thì điểm tối ưu cho sự bùng nổ trí tuệ sẽ là khi tri hoãn nó trong một năm nữa sẽ tạo ra ít hơn hơn $x + y \%$ giảm rủi ro diệt vong liên quan đến sự bùng nổ trí tuệ.

32. Tôi mang ơn Carl Shulman và Stuart Armstrong vì đã giúp đỡ với mô hình này. Xem thêm Shulman (2010a, 3): “Chalmer (2010) đã ghi nhận sự đồng thuận giữa các học viên và nhân viên tại học viện quân sự West Point của Mỹ rằng chính phủ Mỹ sẽ không kiềm chế nghiên cứu AI ngay cả khi đối mặt với thảm họa tiềm tàng, vì sợ rằng các thế lực đối địch sẽ đạt được lợi thế quyết định.

33. Nghĩa là thông tin trong mô hình luôn luôn xấu theo dự đoán. Tất nhiên, tùy thuộc vào thông tin thực sự là gì, trong một số trường hợp, việc thông tin được biết rõ sẽ là tốt, đặc biệt là nếu khoảng cách giữa nhóm dẫn đầu và nhóm bám đuôi hơn nhiều so với những gì người ta có thể đoán trước.

34. Nó thậm chí có thể gây ra rủi ro diệt vong, đặc biệt nếu theo sau sự ra đời của các công nghệ quân sự huỷ diệt mới hoặc các hoạt động tích lũy vũ khí chưa từng có.

35. Một dự án có thể có nhân viên phân phối trên một số lượng lớn các địa điểm và hợp tác thông qua các kênh truyền thông được mã hóa. Nhưng chiến thuật này liên quan đến sự đánh đổi an ninh: trong khi phân tán địa lý có thể cung cấp một số biện pháp bảo vệ chống lại các cuộc tấn công quân sự, nó sẽ gây trở ngại cho hoạt động an ninh, vì khó ngăn chặn nhân viên đào thoát, làm rò rỉ thông

tin hoặc bị bắt cóc bởi một thế lực đối địch nếu họ phân tán trên quá nhiều địa điểm.

36. Lưu ý rằng một hệ số giảm theo thời gian lớn có thể khiến dự án hoạt động như thể nó đang trong một cuộc đua, ngay cả khi biết rằng nó không có đối thủ cạnh tranh thực sự. Hệ số giảm lớn có nghĩa là nó sẽ quan tâm rất ít đến tương lai xa. Tùy thuộc vào tình huống, điều này sẽ ngăn cản các hoạt động R&D xa vời, có xu hướng trì hoãn cuộc cách mạng trí tuệ máy (mặc dù có thể làm cho nó xảy ra đột ngột hơn, vì có mái hiên phần cứng). Nhưng hệ số giảm lớn (hay mức độ quan tâm thấp cho các thế hệ tương lai) cũng sẽ khiến cho những rủi ro diệt vong dường như ít quan trọng hơn. Điều này sẽ khuyến khích các trò may rủi liên quan đến khả năng thu lợi tức thì với cái giá phải trả là tăng nguy cơ xảy ra thảm họa diệt vong, và vì thế sẽ kìm hãm các khoản đầu tư vào an toàn và thúc đẩy phát động sớm sự bùng nổ (giống như hiệu ứng của động lực chạy đua). Tuy nhiên, trái ngược với động lực chạy đua, một hệ số giảm lớn (hoặc không quan tâm đến các thế hệ tương lai) sẽ không có xu hướng đặc biệt để kích động xung đột.

Giảm năng động lực chạy đua là một lợi ích chính của sự hợp tác. Một lợi ích khác là sự hợp tác sẽ tạo điều kiện chia sẻ ý tưởng về cách giải quyết vấn đề kiểm soát, mặc dù điều này ở một mức độ nào đó đối trọng với thực tế là việc hợp tác cũng sẽ tạo điều kiện chia sẻ ý tưởng về cách giải quyết vấn đề năng lực. Hiệu quả của việc tạo điều kiện chia sẻ ý tưởng này có thể làm tăng nhẹ trí tuệ tập thể của cộng đồng nghiên cứu có liên quan.

37. Mặt khác, sự giám sát công khai của một chính phủ duy nhất sẽ có nguy cơ tạo ra kết quả trong đó một quốc gia độc quyền thu lợi.

Kết quả này có vẻ bất lợi hơn kết quả mà trong đó sự vị tha sẽ đảm bảo rằng mọi người đều có thể đạt được lợi ích. Hơn nữa, sự giám sát từ chính phủ của một quốc gia không hẳn đồng nghĩa với việc toàn bộ công dân của quốc gia đó cũng nhận được một phần lợi ích: tùy thuộc vào quốc gia được đề cập, rủi ro của việc toàn bộ các lợi ích sẽ bị thôn tóm bởi một người quyền cao chức trọng trong giới chính trị hoặc các cá nhân vị kỷ.

38. Một điều kiện là việc sử dụng gói phần thưởng (như đã thảo luận trong Chương 12) trong một số trường hợp có thể khuyến khích mọi người tham gia dự án với tư cách là cộng tác viên tích cực thay vì những người tự do thụ động.

39. Hiệu suất giảm dần dường như được đặt ở quy mô nhỏ hơn nhiều. Hầu hết mọi người thà có một ngôi sao hơn là cơ hội một phần tỷ của việc có một thiên hà với một tỷ ngôi sao. Thật vậy, hầu hết mọi người thà có một phần tỷ tài nguyên trên Trái đất hơn là cơ hội một phần tỷ của việc sở hữu toàn bộ hành tinh.

40. So sánh với Shulman (2010a).

41. Các lý thuyết đạo đức tổng hợp gặp rắc rối khi ý tưởng rằng vũ trụ có thể là vô hạn được xem xét nghiêm túc; xem Bostrom (2011b). Cũng có thể có rắc rối khi ý tưởng về các giá trị lớn nhưng hữu hạn được xem xét nghiêm túc; xem Bostrom (2009b).

42. Nếu làm cho máy tính lớn hơn, cuối cùng người ta sẽ phải đối mặt với những hạn chế tương đối phát sinh từ độ trễ giao tiếp giữa các phần khác nhau của tín hiệu máy tính vốn không truyền nhanh hơn ánh sáng. Nếu thu nhỏ máy tính, người ta sẽ gặp phải các giới hạn lượng tử của việc thu nhỏ. Nếu tăng mật độ của máy tính, người ta sẽ rơi vào giới hạn lỗ đen. Phải thừa nhận, chúng ta không thể

hoàn toàn chắc chắn rằng một ngành vật lý mới sẽ không được phát hiện một ngày nào đó và đưa ra một số cách giải quyết những hạn chế này.

43. Số lượng bản sao của một người sẽ tăng tuyến tính với các tài nguyên mà không có mức trần. Tuy nhiên, việc con người bình thường coi trọng việc có nhiều bản sao của mình đến đâu là chưa rõ ràng. Ngay cả những người muốn được sao chép lại nhiều lần cũng có thể không có hàm thỏa dụng là tuyến tính với số lượng bản sao tăng lên. Các con số sao chép, chẳng hạn như tuổi thọ, có thể có hiệu suất giảm dần trong hàm thỏa dụng đặc thù của con người.

44. Một thể đơn nhất có mức độ hợp tác nội bộ rất cao khi ra quyết định.

Một thể đơn nhất có thể có mức độ không hợp tác và xung đột cao hơn ở các cấp thấp hơn, nếu tác tử cấp cao hơn cấu thành nên thể đơn nhất lựa chọn việc mọi thứ xảy ra theo cách đó.

45. Nếu mỗi nhóm AI đối địch bị thuyết phục rằng các nhóm khác bị lạc đường đến mức không có cơ hội tạo ra sự bùng nổ trí tuệ, một lý do cho sự hợp tác – tránh khỏi động lực chạy đua – sẽ bị loại bỏ: mỗi nhóm nên độc lập chọn đi chậm hơn với sự tự tin rằng mình không hề bị cạnh tranh một cách quyết liệt.

46. Một nghiên cứu sinh tiến sĩ.

47. Phát biểu này dự định được hiểu là bao chứa một quy định rằng sự “khỏe mạnh” của các loài không phải con người và các thực thể có tri giác khác (bao gồm cả tâm trí kỹ thuật số) đang tồn tại hoặc có thể sẽ tồn tại được xem xét thích hợp. Nó không phải là giấy phép cho một nhà phát triển AI để thay thế trực giác đạo đức của chính mình với trực giác đạo đức của những người thuộc cộng đồng

đạo đức rộng lớn hơn. Nguyên tắc này phù hợp với phương pháp tiếp cận CEV của nhân loại được thảo luận trong Chương 12, với một cơ sở ngoại suy bao gồm tất cả mọi người.

Để làm rõ thêm: Phát biểu này không nhất thiết loại trừ khả năng của các quyền sở hữu sau chuyển đổi đối với các siêu trí tuệ nhân tạo hoặc các thuật toán cấu thành và cấu trúc dữ liệu của chúng. Phát biểu này không để tâm đến việc những hệ thống pháp lý hoặc chính trị nào sẽ phù hợp nhất để tổ chức các giao dịch trong một xã hội hậu nhân giả thuyết tương lai. Điều mà phát biểu này muốn khẳng định là lựa chọn của một hệ thống như vậy, trong chừng mực mà sự lựa chọn của nó được xác định bởi cách thức siêu trí tuệ được phát triển ban đầu, nên được thực hiện trên cơ sở tiêu chí đã nêu; nghĩa là, hệ thống thể chế hậu chuyển đổi nên được lựa chọn vì lợi ích của toàn nhân loại và phục vụ cho những lý tưởng đạo đức được chia sẻ rộng rãi, ví dụ như, vì lợi ích của bất kỳ ai là người đầu tiên tình cờ phát triển siêu trí tuệ.

48. Điều khoản lợi nhuận bất ngờ hoàn toàn có thể được tinh chỉnh. Ví dụ: có lẽ ngưỡng phải được biểu thị theo bình quân đầu người, hoặc có thể nhóm chiến thắng nên được phép giữ một phần lớn hơn so với mức bình quân của khoản thắng dư để khuyến khích việc sản xuất thêm một cách mạnh mẽ hơn (một phiên bản nào đó của nguyên tắc maximin của Rawls có thể sẽ hấp dẫn ở đây). Các phiên bản tinh chỉnh khác sẽ hướng sự quan tâm của điều khoản ra khỏi vấn đề tiền bạc và phát biểu lại nó trên khía cạnh “tầm ảnh hưởng đến tương lai nhân loại” hoặc “mức độ mà ở đó lợi ích của các bên được đánh trọng số trong hàm thỏa dụng của một thể đơn nhất tương lai”, hoặc tương tự.

CHƯƠNG 15: GIAI ĐOẠN NƯỚC RÚT

1. Một số nghiên cứu đáng giá không phải vì những gì nó khám phá mà vì những lý do khác, chẳng hạn như bằng cách giải trí, giáo dục, công nhận hoặc nâng đỡ những người tham gia vào nó.

2. Tôi không đề xuất rằng không một ai nên nghiên cứu toán học hay triết học thuần túy, và cũng không gợi ý rằng những nỗ lực này đặc biệt lãng phí so với tất cả mọi hoạt động gây hao tán khác trong giới hàn lâm cũng như xã hội nói chung. Có lẽ rất tốt khi một số người có thể cống hiến cho lối sống tư duy và đi theo trí tò mò của mình đến bất cứ đâu, không phụ thuộc vào bất kỳ suy nghĩ nào về độ thỏa dụng hoặc tác động. Lời đề xuất ở đây là trong giai đoạn kế tiếp, một số bộ óc tốt nhất có thể, khi nhận ra rằng hiệu năng nhận thức của mình có thể trở nên lỗi thời trong tương lai gần, muốn chuyển sự chú ý của họ sang những vấn đề lý thuyết mà ở đó, việc tìm ra giải pháp sớm hơn dù chỉ tích tắc cũng sẽ đem đến khác biệt.

3. Mặc dù vậy, người ta nên thận trọng trong trường hợp sự không chắc chắn này có thể mang tính phòng vệ. Chẳng hạn, hãy nhớ lại mô hình cuộc chạy đua rủi ro trong Khung 13 mà ở đó, ta thấy rằng thông tin chiến lược bổ sung có thể gây hại. Tổng quát hơn, chúng ta cần lo lắng về các mối nguy hại về thông tin (xem Bostrom [2011b]). Thật hấp dẫn khi nói rằng chúng ta cần phân tích thêm các mối nguy hại thông tin. Điều này có lẽ đúng, mặc dù chúng ta vẫn có thể lo lắng rằng chính việc phân tích đó có thể tạo ra thông tin nguy hiểm.

4. So sánh với Bostrom (2007).

5. Tôi biết ơn Carl Shulman vì đã nhấn mạnh điểm này.

Tài liệu tham khảo

Acemoglu, Daron. 2003. "Labor- and Capital-Augmenting Technical Change." *Journal of the European Economic Association* 1 (1): 1-37.

Albertson, D. G., and Thomson, J. N. 1976. "The Pharynx of *Caenorhabditis Elegans*." *Philosophical Transactions of the Royal Society B: Biological Sciences* 275 (938): 299-325. Allen, Robert C. 2008. "A Review of Gregory Clark's *A Farewell to Alms: A Brief Economic History of the World*." *Journal of Economic Literature* 46 (4): 946-73.

American Horse Council. 2005. "National Economic Impact of the US Horse Industry." Retrieved July 30, 2013. Available at <http://www.horsecouncil.org/national-economic-impact-us-horse-industry>.

Anand, Paul, Pattanaik, Prasanta, and Puppe, Clemens, eds. 2009. *The Oxford Handbook of Rational and Social Choice*. New York: Oxford University Press.

Andres, B., Koethe, U., Kroeger, T., Helmstaedter, M., Briggman, K. L., Denk, W., and Hamprecht, F. A. 2012. "3D Segmentation of SBFSEM images of Neuropil by a Graphical Model over Supervoxel Boundaries." *Medical Image Analysis* 16 (4): 796-805.

Armstrong, Alex. 2012. "Computer Competes in Crossword Tournament." *I Programmer*, March 19.

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Unpublished manuscript, October 20. Retrieved December 31, 2012. Available at <http://www.neweuropeancentury.org/GodAI.pdf>.

Armstrong, Stuart. 2010. Utility Indifference, Technical Report 2010-1. Oxford: Future of Humanity Institute, University of Oxford.

Armstrong, Stuart. 2013. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* 12: 68-84.

Armstrong, Stuart, and Sandberg, Anders. 2013. "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox." *Acta Astronautica* 89: 1-13.

Armstrong, Stuart, and Sotala, Kaj. 2012. "How We're Predicting AI—or Failing To." In *Beyond AI: Artificial Dreams*, edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52-75. Pilsen: University of West Bohemia. Retrieved February 2, 2013.

Asimov, Isaac. 1942. "Runaround." *Astounding Science-Fiction*, March, 94-103. Asimov, Isaac. 1985. *Robots and Empire*. New York: Doubleday.

Aumann, Robert J. 1976. "Agreeing to Disagree." *Annals of Statistics* 4 (6): 1236-9.

Averch, Harvey Allen. 1985. *A Strategic Analysis of Science and Technology Policy*. Baltimore: Johns Hopkins University Press.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Jacob, W., Lent, R., and Herculano-Houzel, S. 2009. "Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain."

Journal of Comparative Neurology 513 (5): 532-41.

Baars, Bernard J. 1997. In the Theater of Consciousness: The Workspace of the Mind. New York: Oxford University Press.

Baratta, Joseph Preston. 2004. The Politics of World Federation: United Nations, UN Reform, Atomic Control. Westport, CT: Praeger.

Barber, E. J. W. 1991. Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean. Princeton, NJ: Princeton University Press.

Bartels, J., Andreasen, D., Ehirim, P., Mao, H., Seibert, S., Wright, E. J., and Kennedy, P. 2008. "Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex." Journal of Neuroscience Methods 174 (2): 168-76.

Bartz, Jennifer A., Zaki, Jamil, Bolger, Niall, and Ochsner, Kevin N. 2011. "Social Effects of Oxytocin in Humans: Context and Person Matter." Trends in Cognitive Science 15 (7): 301-9.

Basten, Stuart, Lutz, Wolfgang, and Scherbov, Sergei. 2013. "Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility." Demographic Research 28: 1145-66.

Baum, Eric B. 2004. What Is Thought? Bradford Books. Cambridge, MA: MIT Press.

Baum, Seth D., Goertzel, Ben, and Goertzel, Ted G. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." Technological Forecasting and Social Change 78 (1): 185-95.

Beal, J., and Winston, P. 2009. "Guest Editors' Introduction: The New Frontier of Human- Level Artificial Intelligence." IEEE Intelligent Systems 24 (4): 21-3.

Bell, C. Gordon, and Gemmell, Jim. 2009. *Total Recall: How the E-Memory Revolution Will Change Everything*. New York: Dutton.

Benyamin, B., Pourcain, B. St., Davis, O. S., Davies, G., Hansell, M. K., Brion, M.-J. A., Kirkpatrick, R. M., et al. 2013. "Childhood Intelligence is Heritable, Highly Polygenic and Associated With FBNP1L." *Molecular Psychiatry* (January 23).

Berg, Joyce E., and Rietz, Thomas A. 2003. "Prediction Markets as Decision Support Systems." *Information Systems Frontiers* 5 (1): 79-93.

Berger, Theodore W., Chapin, J. K., Gerhardt, G. A., Soussou, W. V., Taylor, D. M., and Tresco, P. A., eds. 2008. *Brain-Computer Interfaces: An International Assessment of Research and Development Trends*. Springer.

Berger, T. W., Song, D., Chan, R. H., Marmarelis, V. Z., LaCoss, J., Wills, J., Hampson, R. E., Deadwyler, S. A., and Granacki, J. J. 2012. "A Hippocampal Cognitive Prosthesis: Multi- Input, Multi-Output Nonlinear Modeling and VLSI Implementation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2): 198-211.

Berliner, Hans J. 1980a. "Backgammon Computer-Program Beats World Champion." *Artificial Intelligence* 14 (2): 205-220.

Berliner, Hans J. 1980b. "Backgammon Program Beats World Champ." *SIGART Newsletter* 69: 6-9.

Bernardo, José M., and Smith, Adrian F. M. 1994. *Bayesian Theory*, 1st ed. Wiley Series in Probability & Statistics. New York: Wiley.

Birbaumer, N., Murguialday, A. R., and Cohen, L. 2008. "Brain-Computer Interface in Paralysis." *Current Opinion in Neurology* 21 (6):

634-8.

Bird, Jon, and Layzell, Paul. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." In *Proceedings of the 2002 Congress on Evolutionary Computation*, 2: 1836-41.

Blair, Clay, Jr. 1957. "Passing of a Great Mind: John von Neumann, a Brilliant, Jovial Mathematician, was a Prodigious Servant of Science and His Country." *Life*, February 25, 89-104.

Bobrow, Daniel G. 1968. "Natural Language Input for a Computer Problem Solving System." In *Semantic Information Processing*, edited by Marvin Minsky, 146 -227. Cambridge, MA: MIT Press.

Bostrom, Nick. 1997. "Predictions from Philosophy? How Philosophers Could Make Themselves Useful." Unpublished manuscript. Last revised September 19, 1998.

Bostrom, Nick. 2002a. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York: Routledge.

Bostrom, Nick. 2002b. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9.

Bostrom, Nick. 2003a. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243-55.

Bostrom, Nick. 2003b. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308-314.

Bostrom, Nick. 2003c. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit and

George E. Lasker, 2: 12-17. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Bostrom, Nick. 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing*, edited by Charles Tandy, 2: 339-371. Death and Anti-Death. Palo Alto, CA: Ria University Press.

Bostrom, Nick. 2006a. "How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5(1): 11-30.

Bostrom, Nick. 2006b. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." *Minds and Machines* 16 (2): 185-200.

Bostrom, Nick. 2006c. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 48-54.

Bostrom, Nick. 2007. "Technological Revolutions: Ethics and Policy in the Dark." In *Nanoscale: Issues and Perspectives for the Nano Century*, edited by Nigel M. de S. Cameron and M. Ellen Mitchell, 129-52. Hoboken, NJ: Wiley.

Bostrom, Nick. 2008a. "Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing." *MIT Technology Review*, May/June issue, 72-7.

Bostrom, Nick. 2008b. "Why I Want to Be a Posthuman When I Grow Up." In *Medical Enhancement and Posthumanity*, edited by Bert Gordijn and Ruth Chadwick, 107-37. New York: Springer.

Bostrom, Nick. 2008c. "Letter from Utopia." *Studies in Ethics, Law, and Technology* 2 (1): 1-7.

Bostrom, Nick. 2009a. "Moral Uncertainty - Towards a Solution?" *Overcoming Bias* (blog), January 1.

Bostrom, Nick. 2009b. "Pascal's Mugging." *Analysis* 69 (3): 443-5.

Bostrom, Nick. 2009c. "The Future of Humanity." In *New Waves in Philosophy of Technology*, edited by Jan Kyrre Berg Olsen, Evan Selinger, and Søren Riis, 186-215. New York: Palgrave Macmillan.

Bostrom, Nick. 2011a. "Information Hazards: A Typology of Potential Harms from Knowledge." *Review of Contemporary Philosophy* 10: 44-79.

Bostrom, Nick. 2011b. "Infinite Ethics." *Analysis and Metaphysics* 10: 9-59.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Muller, special issue, *Minds and Machines* 22 (2): 71-85.

Bostrom, Nick, and Cirkovic, Milan M. 2003. "The Doomsday Argument and the Self-Indication Assumption: Reply to Olum." *Philosophical Quarterly* 53 (210): 83-91.

Bostrom, Nick, and Ord, Toby. 2006. "The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics." *Ethics* 116 (4): 656-79.

Bostrom, Nick, and Roache, Rebecca. 2011. "Smart Policy: Cognitive Enhancement and the Public Interest." In *Enhancing Human Capacities*, edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane, 138-49. Malden, MA: Wiley-Blackwell.

Bostrom, Nick and Sandberg, Anders. 2009a. "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges." *Science and Engineering Ethics* 15 (3): 311-41.

Bostrom, Nick and Sandberg, Anders. 2009b. "The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement." In *Human Enhancement*, 1st ed., edited by Julian Savulescu and Nick Bostrom, 375-416. New York: Oxford University Press.

Bostrom, Nick, Sandberg, Anders, and Douglas, Tom. 2013. "The Unilateralist's Curse: The Case for a Principle of Conformity." Working Paper. Retrieved February 28, 2013. Available at <http://www.nickbostrom.com/papers/unilateralist.pdf>.

Bostrom, Nick, and Yudkowsky, Eliezer. Forthcoming. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.

Boswell, James. 1917. *Boswell's Life of Johnson*. New York: Oxford University Press. Bouchard, T. J. 2004. "Genetic Influence on Human Psychological Traits: A Survey." *Current Directions in Psychological Science* 13 (4): 148-51.

Bourget, David, and Chalmers, David. 2009. "The PhilPapers Surveys." November. Available at <http://philpapers.org/surveys/>.

Bradbury, Robert J. 1999. "Matrioshka Brains." Archived version. As revised August 16, 2004. Available at <http://web.archive.org/web/20090615040912/http://www.aeiveos.com/~bradbury/MatrioshkaBrains/MatrioshkaBrainsPaper.html>.

Brinton, Crane. 1965. *The Anatomy of Revolution*. Revised ed. New York: Vintage Books. Bryson, Arthur E., Jr., and Ho, Yu-Chi. 1969. *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham, MA: Blaisdell.

Buehler, Martin, Iagnemma, Karl, and Singh, Sanjiv, eds. 2009. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*. Springer Tracts in Advanced Robotics 56. Berlin: Springer.

Burch-Brown, J. 2014. "Clues for Consequentialists." *Utilitas* 26 (1): 105-19.

Burke, Colin. 2001. "Agnes Meyer Driscoll vs. the Enigma and the Bombe." Unpublished manuscript. Retrieved February 22, 2013. Available at <http://userpages.umbc.edu/~burke/driscoll1-2011.pdf>.

Canback, S., Samouel, P., and Price, D. 2006. "Do Diseconomies of Scale Impact Firm Size and Performance? A Theoretical and Empirical Overview." *Journal of Managerial Economics* 4 (1): 27-70.

Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov, D. F., Patil, P. G., Henriquez, C. S., and Nicolelis, M. A. 2003. "Learning to Control a Brain-Machine Interface for Reaching and Grasping by Primates." *Public Library of Science Biology* 1 (2): 193-208.

Carroll, Bradley W., and Ostlie, Dale A. 2007. *An Introduction to Modern Astrophysics*. 2nd ed. San Francisco: Pearson Addison Wesley.

Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Carter, Brandon. 1983. "The Anthropic Principle and its Implications for Biological Evolution." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 310 (1512): 347-63.

Carter, Brandon. 1993. "The Anthropic Selection Principle and the Ultra-Darwinian Synthesis." In *The Anthropic Principle: Proceedings*

of the Second Venice Conference on Cosmology and Philosophy, edited by F. Bertola and U. Curi, 33-66. Cambridge: Cambridge University Press.

CFTC & SEC (Commodity Futures Trading Commission and Securities & Exchange Commission). 2010. Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. Washington, DC.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9-10): 7-65.

Chason, R. J., Csokmay, J., Segars, J. H., DeCherney, A. H., and Armant, D. R. 2011. "Environmental and Epigenetic Effects Upon Preimplantation Embryo Metabolism and Development." *Trends in Endocrinology and Metabolism* 22 (10): 412-20.

Chen, S., and Ravallion, M. 2010. "The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty." *Quarterly Journal of Economics* 125 (4): 1577-1625.

Chislenko, Alexander. 1996. "Networking in the Mind Age: Some Thoughts on Evolution of Robotics and Distributed Systems." Unpublished manuscript.

Chislenko, Alexander. 1997. "Technology as Extension of Human Functional Architecture." *Extropy Online*.

Chorost, Michael. 2005. *Rebuilt: How Becoming Part Computer Made Me More Human*. Boston: Houghton Mifflin.

Christiano, Paul F. 2012. "'Indirect Normativity' Write-up." *Ordinary Ideas* (blog), April 21. CIA. 2013. "The World Factbook." Central Intelligence Agency. Retrieved August 3. Available at

[https://www.cia.gov/library/publications/the-world-factbook/rankorder/2127rank.html?](https://www.cia.gov/library/publications/the-world-factbook/rankorder/2127rank.html?countryname=United%20States&countrycode=us®ionCode=na&rank=121#us)

countryname=United%20States&countrycode=us®ionCode=na&rank=121#us. Cicero. 1923. "On Divination." In *On Old Age, on Friendship, on Divination*, translated by

W. A. Falconer. Loeb Classical Library. Cambridge, MA: Harvard University Press. Cirasella, Jill, and Kopec, Danny. 2006. "The History of Computer Games." Exhibit at Dartmouth Artificial Intelligence Conference: The Next Fifty Years (AI@50), Dartmouth College, July 13-15.

Cirkovic, Milan M. 2004. "Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings." *Foundations of Physics* 34 (2): 239-61.

Cirkovic, Milan M., Sandberg, Anders, and Bostrom, Nick. 2010. "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks." *Risk Analysis* 30 (10): 1495-1506.

Clark, Andy, and Chalmers, David J. 1998. "The Extended Mind." *Analysis* 58 (1): 7-19. Clark, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World*. 1st ed.

Princeton, NJ: Princeton University Press.

Clavin, Whitney. 2012. "Study Shows Our Galaxy Has at Least 100 Billion Planets." Jet Propulsion Laboratory, January 11.

CME Group. 2010. What Happened on May 6th? Chicago, May 10.

Coase, R. H. 1937. "The Nature of the Firm." *Economica* 4 (16): 386-405.

Cochran, Gregory, and Harpending, Henry. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York:

Basic Books.

Cochran, G., Hardy, J., and Harpending, H. 2006. "Natural History of Ashkenazi Intelligence." *Journal of Biosocial Science* 38 (5): 659-93.

Cook, James Gordon. 1984. *Handbook of Textile Fibres: Natural Fibres*. Cambridge: Woodhead.

Cope, David. 1996. *Experiments in Musical Intelligence*. Computer Music and Digital Audio Series. Madison, WI: A-R Editions.

Cotman, Carl W., and Berchtold, Nicole C. 2002. "Exercise: A Behavioral Intervention to Enhance Brain Health and Plasticity." *Trends in Neurosciences* 25 (6): 295-301.

Cowan, Nelson. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24 (1): 87-114.

Crabtree, Steve. 1999. "New Poll Gauges Americans' General Knowledge Levels." *Gallup News*, July 6.

Cross, Stephen E., and Walker, Edward. 1994. "Dart: Applying Knowledge Based Planning and Scheduling to Crisis Action Planning." In *Intelligent Scheduling*, edited by Monte Zweben and Mark Fox, 711-29. San Francisco, CA: Morgan Kaufmann.

Crow, James F. 2000. "The Origins, Patterns and Implications of Human Spontaneous Mutation." *Nature Reviews Genetics* 1 (1): 40-7.

Cyranoski, David. 2013. "Stem Cells: Egg Engineers." *Nature* 500 (7463): 392-4.

Dagnelie, Gislin. 2012. "Retinal Implants: Emergence of a Multidisciplinary Field." *Current Opinion in Neurology* 25 (1): 67-75.

Dai, Wei. 2009. "Towards a New Decision Theory." Less Wrong (blog), August 13. Dalrymple, David. 2011. "Comment on Kaufman, J. 'Whole Brain Emulation: Looking at Progress on C. Elegans.'" Less Wrong (blog), October 29.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., Ke, X., et al. 2011. "GenomeWide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic." *Molecular Psychiatry* 16 (10): 996-1005.

Davis, Oliver S. P., Butcher, Lee M., Docherty, Sophia J., Meaburn, Emma L., Curtis, Charles J. C., Simpson, Michael A., Schalkwyk, Leonard C., and Plomin, Robert. 2010. "A Three- Stage Genome-Wide Association Study of General Cognitive Ability: Hunting the Small Effects." *Behavior Genetics* 40 (6): 759-767.

Dawkins, Richard. 1995. *River Out of Eden: A Darwinian View of Life*. Science Masters Series. New York: Basic Books.

De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems*. Machine Intelligence Research Institute, San Francisco, CA, May 19.

De Long, J. Bradford. 1998. "Estimates of World GDP, One Million B.C.-Present." Unpublished manuscript.

De Raedt, Luc, and Flach, Peter, eds. 2001. *Machine Learning: ECML 2001: 12th European Conference on Machine Learning*, Freiburg, Germany, September 5-7, 2001. Proceedings. Lecture Notes in Computer Science 2167. New York: Springer.

Dean, Cornelia. 2005. "Scientific Savvy? In U.S., Not Much." *New York Times*, August 30. Deary, Ian J. 2001. "Human Intelligence

Differences: A Recent History.” *Trends in Cognitive Sciences* 5 (3): 127-30.

Deary, Ian J. 2012. “Intelligence.” *Annual Review of Psychology* 63: 453-82.

Deary, Ian J., Penke, L., and Johnson, W. 2010. “The Neuroscience of Human Intelligence Differences.” *Nature Reviews Neuroscience* 11 (3): 201-11.

Degnan, G. G., Wind, T. C., Jones, E. V., and Edlich, R. F. 2002. “Functional Electrical Stimulation in Tetraplegic Patients to Restore Hand Function.” *Journal of Long-Term Effects of Medical Implants* 12 (3): 175-88.

Devlin, B., Daniels, M., and Roeder, K. 1997. “The Heritability of IQ.” *Nature* 388 (6641):468-71.

Dewey, Daniel. 2011. “Learning What to Value.” In *Artificial General Intelligence: 4th International Conference, AGI2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, edited by Jurgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 309-14. *Lecture Notes in Computer Science* 6830. Berlin: Springer.

Dowe, D. L., and Hernández-Orallo, J. 2012. “IQ Tests Are Not for Machines, Yet.” *Intelligence* 40 (2): 77-81.

Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics*. Bradford Books. Cambridge, MA: MIT Press.

Drexler, K. Eric. 1986. *Engines of Creation*. Garden City, NY: Anchor.

Drexler, K. Eric. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. New York: Wiley.

Drexler, K. Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*. New York: PublicAffairs.

Driscoll, Kevin. 2012. "Code Critique: 'Altair Music of a Sort.'" Paper presented at Critical Code Studies Working Group Online Conference, 2012, February 6.

Dyson, Freeman J. 1960. "Search for Artificial Stellar Sources of Infrared Radiation." *Science* 131 (3414): 1667-1668.

Dyson, Freeman J. 1979. *Disturbing the Universe*. 1st ed. Sloan Foundation Science Series. New York: Harper & Row.

Elga, Adam. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69 (2): 383-96.

Elga, Adam. 2007. "Reflection and Disagreement." *Nous* 41 (3): 478-502.

Eliasmith, Chris, Stewart, Terrence C., Choo, Xuan, Bekolay, Trevor, DeWolf, Travis, Tang, Yichuan, and Rasmussen, Daniel. 2012. "A Large-Scale Model of the Functioning Brain." *Science* 338(6111): 1202-5.

Ellis, J. H. 1999. "The History of Non-Secret Encryption." *Cryptologia* 23 (3): 267-73. Elyasaf, Achiya, Hauptmann, Ami, and Sipper, Moche. 2011. "Ga-Freecell: Evolving Solvers for the Game of Freecell." In *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference, 1931-1938. GECCO' 11*. New York: ACM.

Eppig, C., Fincher, C. L., and Thornhill, R. 2010. "Parasite Prevalence and the Worldwide Distribution of Cognitive Ability." *Proceedings of the Royal Society B: Biological Sciences* 277 (1701): 3801-8.

Espenshade, T. J., Guzman, J. C., and Westoff, C. F. 2003. "The Surprising Global Variation in Replacement Fertility." *Population Research and Policy Review* 22 (5-6): 575-83.

Evans, Thomas G. 1964. "A Heuristic Program to Solve Geometric-Analogy Problems." In *Proceedings of the April 21-23, 1964, Spring Joint Computer Conference*, 327-338. AFIPS '64. New York: ACM.

Evans, Thomas G. 1968. "A Program for the Solution of a Class of Geometric-Analogy Intelligence Test Questions." In *Semantic Information Processing*, edited by Marvin Minsky, 271-353. Cambridge, MA: MIT Press.

Faisal, A. A., Selen, L. P., and Wolpert, D. M. 2008. "Noise in the Nervous System." *Nature Reviews Neuroscience* 9 (4): 292-303.

Faisal, A. A., White, J. A., and Laughlin, S. B. 2005. "Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring." *Current Biology* 15 (12): 1143-9.

Feldman, Jacob. 2000. "Minimization of Boolean Complexity in Human Concept Learning." *Nature* 407 (6804): 630-3.

Feldman, J. A., and Ballard, Dana H. 1982. "Connectionist Models and Their Properties." *Cognitive Science* 6 (3): 205-254.

Foley, J. A., Monfreda, C., Ramankutty, N., and Zaks, D. 2007. "Our Share of the Planetary Pie." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12585-6.

Forgas, Joseph P., Cooper, Joel, and Crano, William D., eds. 2010. *The Psychology of Attitudes and Attitude Change*. Sydney Symposium of Social Psychology. New York: Psychology Press.

Frank, Robert H. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess*. New York: Free Press.

Fredriksen, Kaja Bonesmo. 2012. *Less Income Inequality and More Growth - Are They Compatible?: Part 6. The Distribution of Wealth*. Technical report, OECD Economics Department Working Papers 929. OECD Publishing.

Freitas, Robert A., Jr. 1980. "A Self-Replicating Interstellar Probe." *Journal of the British Interplanetary Society* 33: 251-64.

Freitas, Robert A., Jr. 2000. "Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations." Foresight Institute. April. Retrieved July 28, 2013. Available at <http://www.foresight.org/nano/Ecophagy.html>.

Freitas, Robert A., Jr., and Merkle, Ralph C. 2004. *Kinematic Self-Replicating Machines*. Georgetown, TX: Landes Bioscience.

Gaddis, John Lewis. 1982. *Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy*. New York: Oxford University Press.

Gammoned.net. 2012. "Snowie." Archived version. Retrieved June 30. Available at <http://web.archive.org/web/20070920191840/http://www.gammoned.com/snowie.html>. Gates, Bill. 1975. "Software Contest Winners Announced." *Computer Notes* 1 (2): 1. Georgieff, Michael K. 2007. "Nutrition and the Developing Brain: Nutrient Priorities and Measurement." *American Journal of Clinical Nutrition* 85 (2): 614S-620S.

Gianaroli, Luca. 2000. "Preimplantation Genetic Diagnosis: Polar Body and Embryo Biopsy." *Supplement, Human Reproduction* 15 (4): 69-75.

Gilovich, Thomas, Griffin, Dale, and Kahneman, Daniel, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Gilster, Paul. 2012. "ESO: Habitable Red Dwarf Planets Abundant." *Centauri Dreams* (blog), March 29.

Goldstone, Jack A. 1980. "Theories of Revolution: The Third Generation." *World Politics* 32 (3): 425-53.

Goldstone, Jack A. 2001. "Towards a Fourth Generation of Revolutionary Theory." *Annual Review of Political Science* 4: 139-87.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers*, edited by Franz L. Alt and Morris Rubinoff, 6: 31-88. New York: Academic Press.

Good, Irving John. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1-4): 67-79.

Good, Irving John. 1976. "Book review of 'The Thinking Computer: Mind Inside Matter'" In *International Journal of Man-Machine Studies* 8: 617-20.

Good, Irving John. 1982. "Ethical Machines." In *Intelligent Systems: Practice and Perspective*, edited by J. E. Hayes, Donald Michie, and Y.-H. Pao, 555-60. *Machine Intelligence* 10. Chichester: Ellis Horwood.

Goodman, Nelson. 1954. *Fact, Fiction, and Forecast*. 1st ed. London: Athlone Press.

Gott, J. R., Juric, M., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N., and Brinkmann, J. 2005. "A Map of the Universe." *Astrophysical Journal* 624 (2): 463-83. Gottfredson, Linda

S. 2002. "G: Highly General and Highly Practical." In *The General Factor of Intelligence: How General Is It?*, edited by Robert J. Sternberg and Elena L. Grigorenko, 331-80. Mahwah, NJ: Lawrence Erlbaum.

Gould, S. J. 1990. *Wonderful Life: The Burgess Shale and the Nature of History*. New York: Norton.

Graham, Gordon. 1997. *The Shape of the Past: A Philosophical Approach to History*. New York: Oxford University Press.

Gray, C. M., and McCormick, D. A. 1996. "Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex." *Science* 274 (5284): 109-13.

Greene, Kate. 2012. "Intel's Tiny Wi-Fi Chip Could Have a Big Impact." *MIT Technology Review*, September 21.

Guizzo, Erico. 2010. "World Robot Population Reaches 8.6 Million." *IEEE Spectrum*, April 14.

Gunn, James E. 1982. *Isaac Asimov: The Foundations of Science Fiction*. Science-Fiction Writers. New York: Oxford University Press.

Haberl, Helmut, Erb, Karl-Heinz, and Krausmann, Fridolin. 2013. "Global Human Appropriation of Net Primary Production (HANPP)." *Encyclopedia of Earth*, September 3. Haberl, H., Erb, K. H., Krausmann, F., Gaube, V., Bondeau, A., Plutzer, C., Gingrich, S., Lucht, W., and Fischer-Kowalski, M. 2007. "Quantifying and Mapping the Human Appropriation of Net Primary Production in Earth's Terrestrial Ecosystems." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12942-7. Hájek, Alan. 2009. "Dutch Book Arguments." In Anand, Pattanaik, and Puppe

2009, 173-95. Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine*. Amherst, NY: Prometheus Books.

Hampson, R. E., Song, D., Chan, R. H., Sweatt, A. J., Riley, M. R., Gerhardt, G. A., Shin, D. C., Marmarelis, V. Z., Berger, T. W., and Deadwyler, S. A. 2012. "A Nonlinear Model for Hippocampal Cognitive Prosthesis: Memory Facilitation by Hippocampal Ensemble Stimulation." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2): 184-97.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2). Hanson, Robin. 1995. "Could Gambling Save Science? Encouraging an Honest Consensus."

Social Epistemology 9 (1): 3-33.

Hanson, Robin. 1998a. "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization." Unpublished manuscript, July 1. Retrieved April 26, 2012. <http://hanson.gmu.edu/filluniv.pdf>.

Hanson, Robin. 1998b. "Economic Growth Given Machine Intelligence." Unpublished manuscript. Retrieved May 15, 2013. Available at <http://hanson.gmu.edu/aigrow.pdf>.

Hanson, Robin. 1998c. "Long-Term Growth as a Sequence of Exponential Modes." Unpublished manuscript. Last revised December 2000. Available at <http://hanson.gmu.edu/longgrow.pdf>.

Hanson, Robin. 1998d. "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions." Unpublished manuscript, September 23. Retrieved August 12, 2012. Available at <http://hanson.gmu.edu/hardstep.pdf>.

Hanson, Robin. 2000. "Shall We Vote on Values, But Bet on Beliefs?" Unpublished manuscript, September. Last revised October 2007. Available at <http://hanson.gmu.edu/futarchy.pdf>. Hanson, Robin. 2006. "Uncommon Priors Require Origin Disputes." *Theory and Decision* 61 (4): 319-328.

Hanson, Robin. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45-50. Hanson, Robin. 2009. "Tiptoe or Dash to Future?" *Overcoming Bias* (blog), December 23. Hanson, Robin. 2012. "Envisioning the Economy, and Society, of Whole Brain Emulations."

Paper presented at the AGI Impacts conference 2012.

Hart, Oliver. 2008. "Economica Coase Lecture Reference Points and the Theory of the Firm." *Economica* 75 (299): 404-11.

Hay, Nicholas James. 2005. "Optimal Agents." B.Sc. thesis, University of Auckland. Hedberg, Sara Reese. 2002. "Dart: Revolutionizing Logistics Planning." *IEEE Intelligent Systems* 17 (3): 81-3.

Helliwell, John, Layard, Richard, and Sachs, Jeffrey. 2012. *World Happiness Report*. The Earth Institute.

Helmstaedter, M., Briggman, K. L., and Denk, W. 2011. "High-Accuracy Neurite Reconstruction for High-Throughput Neuroanatomy." *Nature Neuroscience* 14 (8): 1081-8.

Heyl, Jeremy S. 2005. "The Long-Term Future of Space Travel." *Physical Review D* 72 (10):1-4.

Hibbard, Bill. 2011. "Measuring Agent Intelligence via Hierarchies of Environments." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, edited by Jurgen Schmidhuber, Kristinn R. Thórisson,

and Moshe Looks, 303-8. Lecture Notes in Computer Science 6830. Berlin: Springer.

Hinke, R. M., Hu, X., Stillman, A. E., Herkle, H., Salmi, R., and Ugurbil, K. 1993. "Functional Magnetic Resonance Imaging of Broca's Area During Internal Speech." *Neuroreport* 4 (6):675-8.

Hinxton Group. 2008. Consensus Statement: Science, Ethics and Policy Challenges of Pluripotent Stem Cell-Derived Gametes. Hinxton, Cambridgeshire, UK, April 11. Available at http://www.hinxtongroup.org/Consensus_HG08_FINAL.pdf.

Hoffman, David E. 2009. *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*. New York: Doubleday.

Hofstadter, Douglas R. (1979) 1999. *Godel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

Holley, Rose. 2009. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs." *D-Lib Magazine* 15 (3-4).

Horton, Sue, Alderman, Harold, and Rivera, Juan A. 2008. *Copenhagen Consensus 2008 Challenge Paper: Hunger and Malnutrition*. Technical report. Copenhagen Consensus Center, May 11.

Howson, Colin, and Urbach, Peter. 1993. *Scientific Reasoning: The Bayesian Approach*. 2nd ed. Chicago: Open Court.

Hsu, Stephen. 2012. "Investigating the Genetic Basis for Intelligence and Other Quantitative Traits." Lecture given at UC Davis Department of Physics Colloquium, Davis, CA, February 13.

Huebner, Bryce. 2008. "Do You See What We See? An Investigation of an Argument Against Collective Representation." *Philosophical Psychology* 21 (1): 91-112.

Huff, C. D., Xing, J., Rogers, A. R., Witherspoon, D., and Jorde, L. B. 2010. "Mobile Elements Reveal Small Population Size in the Ancient Ancestors of Homo Sapiens." *Proceedings of the National Academy of Sciences of the United States of America* 107 (5): 2147-52.

Huffman, W. Cary, and Pless, Vera. 2003. *Fundamentals of Error-Correcting Codes*. New York: Cambridge University Press.

Hunt, Patrick. 2011. "Late Roman Silk: Smuggling and Espionage in the 6th Century CE." *Philolog*, Stanford University (blog), August 2.

Hutter, Marcus. 2001. "Towards a Universal Theory of Artificial Intelligence Based on Algorithmic Probability and Sequential Decisions." In De Raedt and Flach 2001, 226-38. Hutter, Marcus. 2005. *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability*. Texts in Theoretical Computer Science. Berlin: Springer.

Iliadou, A. N., Janson, P. C., and Cnattingius, S. 2011. "Epigenetics and Assisted Reproductive Technology." *Journal of Internal Medicine* 270 (5): 414-20.

Isaksson, Anders. 2007. *Productivity and Aggregate Growth: A Global Picture*. Technical report 05/2007. Vienna, Austria: UNIDO (United Nations Industrial Development Organization) Research and Statistics Branch.

Jones, Garret. 2009. "Artificial Intelligence and Economic Growth: A Few Finger-Exercises." Unpublished manuscript, January.

Retrieved November 5, 2012. Available at <http://mason.gmu.edu/~gjonesb/AlandGrowth>.

Jones, Vincent C. 1985. *Manhattan: The Army and the Atomic Bomb*. United States Army in World War II. Washington, DC: Center of Military History.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction and Decision Theory. New York: Cambridge University Press.

Judd, K. L., Schmedders, K., and Yeltekin, S. 2012. "Optimal Rules for Patent Races." *International Economic Review* 53 (1): 23-52.

Kalfoglou, A., Suthers, K., Scott, J., and Hudson, K. 2004. *Reproductive Genetic Testing: What America Thinks*. Genetics and Public Policy Center.

Kamm, Frances M. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford Ethics Series. New York: Oxford University Press.

Kandel, Eric R., Schwartz, James H., and Jessell, Thomas M., eds. 2000. *Principles of Neural Science*. 4th ed. New York: McGraw-Hill.

Kansa, Eric. 2003. "Social Complexity and Flamboyant Display in Competition: More Thoughts on the Fermi Paradox." Unpublished manuscript, archived version.

Karnofsky, Holden. 2012. "Comment on 'Reply to Holden on Tool AI.'" *Less Wrong* (blog), August 1. Kasparov, Garry. 1996. "The Day That I Sensed a New Kind of Intelligence." *Time*, March 25, no. 13. Kaufman, Jeff. 2011. "Whole Brain Emulation and Nematodes." *Jeff Kaufman's Blog* (blog), November 2.

Keim, G. A., Shazeer, N. M., Littman, M. L., Agarwal, S., Cheves, C. M., Fitzgerald, J., Grosland, J., Jiang, F., Pollard, S., and Weinmeister, K. 1999. "Proverb: The Probabilistic Cruciverbalist." In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 710-17. Menlo Park, CA: AAAI Press.

Kell, Harrison J., Lubinski, David, and Benbow, Camilla P. 2013. "Who Rises to the Top? Early Indicators." *Psychological Science* 24 (5): 648-59.

Keller, Wolfgang. 2004. "International Technology Diffusion." *Journal of Economic Literature* 42 (3): 752-82.

KGS Go Server. 2012. "KGS Game Archives: Games of KGS player zen19." Retrieved July 22, 2013. Available at <http://www.gokgs.com/gameArchives.jsp?user=zen19d&oldAccounts=t&year=2012&month=3>.

Knill, Emanuel, Laflamme, Raymond, and Viola, Lorenza. 2000. "Theory of Quantum Error Correction for General Noise." *Physical Review Letters* 84 (11): 2525-8.

Koch, K., McLean, J., Segev, R., Freed, M. A., Berry, M. J., Balasubramanian, V., and Sterling, P. 2006. "How Much the Eye Tells the Brain." *Current Biology* 16 (14): 1428-34.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., et al. 2012. "Rate of De Novo Mutations and the Importance of Father's Age to Disease Risk." *Nature* 488: 471-5.

Koomey, Jonathan G. 2011. *Growth in Data Center Electricity Use 2005 to 2010*. Technical report, 08/01/2011. Oakland, CA: Analytics Press.

Koubi, Vally. 1999. "Military Technology Races." *International Organization* 53 (3): 537-65. Koubi, Vally, and Lalman, David. 2007. "Distribution of Power and Military R&D." *Journal of Theoretical Politics* 19 (2): 133-52.

Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., and Lanza, G. 2003. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. 2nd ed. Genetic Programming. Norwell, MA: Kluwer Academic.

Kremer, Michael. 1993. "Population Growth and Technological Change: One Million B.C. to 1990." *Quarterly Journal of Economics* 108 (3): 681-716.

Kruel, Alexander. 2011. "Interview Series on Risks from AI." *Less Wrong Wiki* (blog). Retrieved Oct 26, 2013. Available at http://wiki.lesswrong.com/wiki/Interview_series_on_risks_from_AI.

Kruel, Alexander. 2012. "Q&A with Experts on Risks From AI #2." *Less Wrong* (blog), January 9.

Krusienski, D. J., and Shih, J. J. 2011. "Control of a Visual Keyboard Using an Electrocorticographic Brain-Computer Interface." *Neurorehabilitation and Neural Repair* 25 (4): 323-31.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. 1st ed. Chicago: University of Chicago Press.

Kuipers, Benjamin. 2012. "An Existing, Ecologically-Successful Genus of Collectively Intelligent Artificial Creatures." Paper presented at the 4th International Conference, ICCCI 2012, Ho Chi Minh City, Vietnam, November 28-30.

Kurzweil, Ray. 2001. "Response to Stephen Hawking." *Kurzweil Accelerating Intelligence*. September 5. Retrieved December 31,

2012. Available at [http://www.kurzweilai.net/ response-to-stephen-hawking](http://www.kurzweilai.net/response-to-stephen-hawking).

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Laffont, Jean-Jacques, and Martimort, David. 2002. *The Theory of Incentives: The Principal- Agent Model*. Princeton, NJ: Princeton University Press.

Lancet, The. 2008. "Iodine Deficiency—Way to Go Yet." *The Lancet* 372 (9633): 88. Landauer, Thomas K. 1986. "How Much Do People Remember? Some Estimates of the Quantity of Learned Information in Long-Term Memory." *Cognitive Science* 10 (4):477-93.

Lebedev, Anastasiya. 2004. "The Man Who Saved the World Finally Recognized." *MosNews*, May 21.

Lebedev, M. A., and Nicolelis, M. A. 2006. "Brain-Machine Interfaces: Past, Present and Future." *Trends in Neuroscience* 29 (9): 536-46.

Legg, Shane. 2008. "Machine Super Intelligence." PhD diss., University of Lugano.

Leigh, E. G., Jr. 2010. "The Group Selection Controversy." *Journal of Evolutionary Biology* 23(1): 6-19.

Lenat, Douglas B. 1982. "Learning Program Helps Win National Fleet Wargame Tournament." *SIGART Newsletter* 79: 16-17.

Lenat, Douglas B. 1983. "EURISKO: A Program that Learns New Heuristics and Domain Concepts." *Artificial Intelligence* 21 (1-2): 61-98.

Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy & Public Affairs* 29 (4): 342-70.

Lerner, Josh. 1997. "An Empirical Exploration of a Technology Race." *RAND Journal of Economics* 28 (2): 228-47.

Leslie, John. 1996. *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.

Lewis, David. 1988. "Desire as Belief." *Mind: A Quarterly Review of Philosophy* 97 (387):323-32.

Li, Ming, and Vitányi, Paul M. B. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. New York: Springer.

Lin, Thomas, Mausam, and Etzioni, Oren. 2012. "Entity Linking at Web Scale." In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web- scale Knowledge Extraction (AKBC-WEKEX '12)*, edited by James Fan, Raphael Hoffman, Aditya Kalyanpur, Sebastian Riedel, Fabian Suchanek, and Partha Pratim Talukdar, 84-88. Madison, WI: Omnipress.

Lloyd, Seth. 2000. "Ultimate Physical Limits to Computation." *Nature* 406 (6799): 1047-54. Louis Harris & Associates. 1969. "Science, Sex, and Morality Survey, study no. 1927." *Life Magazine* (New York) 4.

Lynch, Michael. 2010. "Rate, Molecular Spectrum, and Consequences of Human Mutation." *Proceedings of the National Academy of Sciences of the United States of America* 107 (3): 961-8.

Lyons, Mark K. 2011. "Deep Brain Stimulation: Current and Future Clinical Applications." *Mayo Clinic Proceedings* 86 (7): 662-72.

MacAskill, William. 2010. "Moral Uncertainty and Intertheoretic Comparisons of Value." BPhil thesis, University of Oxford.

McCarthy, John. 2007. "From Here to Human-Level AI." *Artificial Intelligence* 171 (18): 1174-82.

McCorduck, Pamela. 1979. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. San Francisco: W. H. Freeman.

Mack, C. A. 2011. "Fifty Years of Moore's Law." *IEEE Transactions on Semiconductor Manufacturing* 24 (2): 202-7.

MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. New York: Cambridge University Press.

McLean, George, and Stewart, Brian. 1979. "Norad False Alarm Causes Uproar." *The National*. Aired November 10. Ottawa, ON: CBC, 2012. News Broadcast.

Maddison, Angus. 1999. "Economic Progress: The Last Half Century in Historical Perspective." In *Facts and Fancies of Human Development: Annual Symposium and Cunningham Lecture, 1999*, edited by Ian Castles. Occasional Paper Series, 1/2000. Academy of the Social Sciences in Australia.

Maddison, Angus. 2001. *The World Economy: A Millennial Perspective*. Development Centre Studies. Paris: Development Centre of the Organisation for Economic Co-operation/Development.

Maddison, Angus. 2005. *Growth and Interaction in the World Economy: The Roots of Modernity*. Washington, DC: AEI Press.

Maddison, Angus. 2007. *Contours of the World Economy, 1-2030 AD: Essays in Macro- Economic History*. New York: Oxford University Press.

Maddison, Angus. 2010. "Statistics of World Population, GDP and Per Capita GDP 1-2008 AD." Retrieved October 26, 2013. Available at

[http://www.ggcd.net/maddison/Historical_ Statistics/ vertical-file_02-2010.xls](http://www.ggcd.net/maddison/Historical_Statistics/vertical-file_02-2010.xls).

Mai, Q., Yu, Y., Li, T., Wang, L., Chen, M. J., Huang, S. Z., Zhou, C., and Zhou, Q. 2007. "Derivation of Human Embryonic Stem Cell Lines from Parthenogenetic Blastocysts." *Cell Research* 17 (12): 1008-19.

Mak, J. N., and Wolpaw, J. R. 2009. "Clinical Applications of Brain-Computer Interfaces: Current State and Future Prospects." *IEEE Reviews in Biomedical Engineering* 2: 187-99.

Mankiw, N. Gregory. 2009. *Macroeconomics*. 7th ed. New York, NY: Worth.

Mardis, Elaine R. 2011. "A Decade's Perspective on DNA Sequencing Technology." *Nature* 470 (7333): 198-203.

Markoff, John. 2011. "Computer Wins on 'Jeopardy!': Trivial, It's Not." *New York Times*, February 16. Markram, Henry. 2006. "The Blue Brain Project." *Nature Reviews Neuroscience* 7 (2): 153-160.

Mason, Heather. 2003. "Gallup Brain: The Birth of In Vitro Fertilization." *Gallup*, August 5. Menzel, Randolph, and Giurfa, Martin. 2001. "Cognitive Architecture of a Mini-Brain: The Honeybee." *Trends in Cognitive Sciences* 5 (2): 62-71.

Metzinger, Thomas. 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Mijic, Roko. 2010. "Bootstrapping Safe AGI Goal Systems." Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8.

Mike, Mike. 2013. "Face of Tomorrow." Retrieved June 30, 2012 . Available at [http:// faceoftomorrow.org](http://faceoftomorrow.org). Milgrom, Paul, and Roberts,

John. 1990. "Bargaining Costs, Influence Costs, and the Organization of Economic Activity." In *Perspectives on Positive Political Economy*, edited by James E. Alt and Kenneth A. Shepsle, 57-89. New York: Cambridge University Press.

Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63 (2): 81-97.

Miller, Geoffrey. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. New York: Doubleday.

Miller, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*. Dallas, TX: BenBella Books.

Minsky, Marvin. 1967. *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: PrenticeHall.

Minsky, Marvin, ed. 1968. *Semantic Information Processing*. Cambridge, MA: MIT Press. Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.'" Unpublished manuscript, October 1. Retrieved December 31, 2012. Available at <http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html>.

Minsky, Marvin. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster.

Minsky, Marvin, and Papert, Seymour. 1969. *Perceptrons: An Introduction to Computational Geometry*. 1st ed. Cambridge, MA: MIT Press.

Moore, Andrew. 2011. "Hedonism." In *The Stanford Encyclopedia of Philosophy*, Winter 2011, edited by Edward N. Zalta. Stanford, CA: Stanford University.

Moravec, Hans P. 1976. "The Role of Raw Power in Intelligence." Unpublished manuscript, May 12. Retrieved August 12, 2012. Available at <http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html>.

Moravec, Hans P. 1980. "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover." PhD diss., Stanford University.

Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

Moravec, Hans P. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1.

Moravec, Hans P. 1999. "Rise of the Robots." *Scientific American*, December, 124-35. Muehlhauser, Luke, and Helm, Louie. 2012. "The Singularity and Machine Ethics." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny S0raker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Muehlhauser, Luke, and Salamon, Anna. 2012. "Intelligence Explosion: Evidence and Import." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny S0raker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Muller, Vincent C., and Bostrom, Nick. Forthcoming. "Future Progress in Artificial Intelligence: A Poll Among Experts." In "Impacts

and Risks of Artificial General Intelligence,” edited by Vincent C. Muller, special issue, *Journal of Experimental and Theoretical Artificial Intelligence*.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

Nachman, Michael W., and Crowell, Susan L. 2000. “Estimate of the Mutation Rate per Nucleotide in Humans.” *Genetics* 156 (1): 297-304.

Nagy, Z. P., and Chang, C. C. 2007. “Artificial Gametes.” *Theriogenology* 67 (1): 99-104. Nagy, Z. P., Kerkis, I., and Chang, C. C. 2008. “Development of Artificial Gametes.”

Reproductive BioMedicine Online 16 (4): 539-44.

NASA. 2013. “International Space Station: Facts and Figures.” Available at http://www.nasa.gov/worldbook/intspacestation_worldbook.html.

Newborn, Monty. 2011. *Beyond Deep Blue: Chess in the Stratosphere*. New York: Springer. Newell, Allen, Shaw, J. C., and Simon, Herbert A. 1958. “Chess-Playing Programs and the

Problem of Complexity.” *IBM Journal of Research and Development* 2 (4): 320-35. Newell, Allen, Shaw, J. C., and Simon, Herbert A. 1959. “Report on a General Problem-

Solving Program: Proceedings of the International Conference on Information Processing.” In *Information Processing*, 256-64. Paris: UNESCO.

Nicolelis, Miguel A. L., and Lebedev, Mikhail A. 2009. “Principles of Neural Ensemble Physiology

Underlying the Operation of Brain-Machine Interfaces.” *Nature Reviews Neuroscience* 10 (7): 530-40.

Nilsson, Nils J. 1984. *Shakey the Robot*, Technical Note 323. Menlo Park, CA: AI Center, SRI International, April.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. New York: Cambridge University Press.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., and Turkheimer, E. 2012. “Intelligence: New Findings and Theoretical Developments.” *American Psychologist* 67 (2): 130-59.

Niven, Larry. 1973. “The Defenseless Dead.” In *Ten Tomorrows*, edited by Roger Elwood, 91-142. New York: Fawcett.

Nordhaus, William D. 2007. “Two Centuries of Productivity Growth in Computing.” *Journal of Economic History* 67 (1): 128-59.

Norton, John D. 2011. “Waiting for Landauer.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42 (3): 184 -98.

Olds, James, and Milner, Peter. 1954. “Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain.” *Journal of Comparative and Physiological Psychology* 47 (6): 419-27.

Olum, Ken D. 2002. “The Doomsday Argument and the Number of Possible Observers.” *Philosophical Quarterly* 52 (207): 164-84.

Omohundro, Stephen M. 2007. “The Nature of Self-Improving Artificial Intelligence.” Paper presented at Singularity Summit 2007, San Francisco, CA, September 8-9.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483-92. *Frontiers in Artificial Intelligence and Applications* 171. Amsterdam: IOS.

Omohundro, Stephen M. 2012. "Rational Artificial Intelligence for the Greater Good." In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny SOraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

O'Neill, Gerard K. 1974. "The Colonization of Space." *Physics Today* 27 (9): 32-40. Oshima, Hideki, and Katayama, Yoichi. 2010. "Neuroethics of Deep Brain Stimulation for Mental Disorders: Brain Stimulation Reward in Humans." *Neurologia medico-chirurgica* 50 (9): 845-52.

Parfit, Derek. 1986. *Reasons and Persons*. New York: Oxford University Press.

Parfit, Derek. 2011. *On What Matters*. 2 vols. The Berkeley Tanner Lectures. New York: Oxford University Press.

Parrington, Alan J. 1997. "Mutually Assured Destruction Revisited." *Airpower Journal* 11 (4). Pasqualotto, Emanuele, Federici, Stefano, and Belardinelli, Marta Olivetti. 2012. "Toward Functioning and Usable Brain-Computer Interfaces (BCIs): A Literature Review." *Disability and Rehabilitation: Assistive Technology* 7 (2): 89-103.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press.

Perlmutter, J. S., and Mink, J. W. 2006. "Deep Brain Stimulation." *Annual Review of Neuroscience* 29: 229-57.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York: Viking.

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control Consortium 2, and Davis, O. S. 2013. "Common DNA Markers Can Account for More than Half of the Genetic Influence on Cognitive Abilities." *Psychological Science* 24 (2): 562-8.

Popper, Nathaniel. 2012. "Flood of Errant Trades Is a Black Eye for Wall Street." *New York Times*, August 1.

Pourret, Olivier, Naim, Patrick, and Marcot, Bruce, eds. 2008. *Bayesian Networks: A Practical Guide to Applications*. Chichester, West Sussex, UK: Wiley.

Powell, A., Shennan, S., and Thomas, M. G. 2009. "Late Pleistocene Demography and the Appearance of Modern Human Behavior." *Science* 324 (5932): 1298-1301.

Price, Huw. 1991. "Agency and Probabilistic Causality." *British Journal for the Philosophy of Science* 42 (2): 157-76.

Qian, M., Wang, D., Watkins, W. E., Gebiski, V., Yan, Y. Q., Li, M., and Chen, Z. P. 2005. "The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China." *Asia Pacific Journal of Clinical Nutrition* 14 (1): 32-42.

Quine, Willard Van Orman, and Ullian, Joseph Silbert. 1978. *The Web of Belief*, ed. Richard Malin Ohmann, vol. 2. New York: Random House.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5-31.

Rajab, Moheeb Abu, Zarfoss, Jay, Monroe, Fabian, and Terzis, Andreas. 2006. "A Multifaceted Approach to Understanding the Botnet Phenomenon." In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, 41-52. New York: ACM.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap.

Read, J. I., and Trentham, Neil. 2005. "The Baryonic Mass Function of Galaxies." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363 (1837): 2693-710.

Repantis, D., Schlattmann, P., Laisney, O., and Heuser, I. 2010. "Modafinil and Methylphenidate for Neuroenhancement in Healthy Individuals: A Systematic Review." *Pharmacological Research* 62 (3): 187-206.

Rhodes, Richard. 1986. *The Making of the Atomic Bomb*. New York: Simon & Schuster. Rhodes, Richard. 2008. *Arsenals of Folly: The Making of the Nuclear Arms Race*. New York: Vintage.

Rietveld, Cornelius A., Medland, Sarah E., Derringer, Jaime, Yang, Jian, Esko, Tonu, Martin, Nicolas W., Westra, Harm-Jan, Shakhbazov, Konstantin, Abdellaoui, Abdel, et al. 2013. "GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment." *Science* 340 (6139): 1467-71.

Ring, Mark, and Orseau, Laurent. 2011. "Delusion, Survival, and Intelligent Agents." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, edited by Jurgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 11-20. *Lecture Notes in Computer Science* 6830. Berlin: Springer.

Ritchie, Graeme, Manurung, Ruli, and Waller, Annalu. 2007. "A Practical Application of Computational Humour." In Proceedings of the 4th International Joint Workshop on Computational Creativity, edited by Amilcar Cardoso and Geraint A. Wiggins, 91-8. London: Goldsmiths, University of London.

Roache, Rebecca. 2008. "Ethics, Speculation, and Values." *NanoEthics* 2 (3): 317-27.

Robles, J. A., Lineweaver, C. H., Grether, D., Flynn, C., Egan, C. A., Pracy, M. B., Holmberg, J., and Gardner, E. 2008. "A Comprehensive Comparison of the Sun to Other Stars: Searching for SelfSelection Effects." *Astrophysical Journal* 684 (1): 691-706.

Roe, Anne. 1953. *The Making of a Scientist*. New York: Dodd, Mead.

Roy, Deb. 2012. "About." Retrieved October 14. Available at <http://web.media.mit.edu/~dkroy/>.

Rubin, Jonathan, and Watson, Ian. 2011. "Computer Poker: A Review." *Artificial Intelligence* 175 (5-6): 958-87.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. "Learning Representations by

Back- Propagating Errors." *Nature* 323 (6088): 533-6.

Russell, Bertrand. 1986. "The Philosophy of Logical Atomism." In *The Philosophy of Logical Atomism and Other Essays 1914-1919*, edited by John G. Slater, 8: 157-244. *The Collected Papers of Bertrand Russell*. Boston: Allen & Unwin.

Russell, Bertrand, and Griffin, Nicholas. 2001. *The Selected Letters of Bertrand Russell: The Public Years, 1914-1970*. New York: Routledge.

Russell, Stuart J., and Norvig, Peter. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Sabrosky, Curtis W. 1952. "How Many Insects Are There?" In *Insects*, edited by United States Department of Agriculture, 1-7. Yearbook of Agriculture. Washington, DC: United States Government Printing Office.

Salamon, Anna. 2009. "When Software Goes Mental: Why Artificial Minds Mean Fast

Endogenous Growth." Working Paper, December 27.

Salem, D. J., and Rowan, A. N. 2001. *The State of the Animals: 2001. Public Policy Series*. Washington, DC: Humane Society Press.

Salverda, W., Nolan, B., and Smeeding, T. M. 2009. *The Oxford Handbook of Economic Inequality*. Oxford: Oxford University Press.

Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210-19.

Sandberg, Anders. 1999. "The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains." *Journal of Evolution and Technology* 5.

Sandberg, Anders. 2010. "An Overview of Models of Technological Singularity." Paper presented at the Roadmaps to AGI and the Future of AGI Workshop, Lugano, Switzerland, March 8.

Sandberg, Anders. 2013. "Feasibility of Whole Brain Emulation." In *Philosophy and Theory of Artificial Intelligence*, edited by Vincent C. Muller, 5: 251-64. *Studies in Applied Philosophy, Epistemology and Rational Ethics*. New York: Springer.

Sandberg, Anders, and Bostrom, Nick. 2006. "Converging Cognitive Enhancements." *Annals of the New York Academy of Sciences* 1093: 201-27.

Sandberg, Anders, and Bostrom, Nick. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report 2008-3. Future of Humanity Institute, University of Oxford.

Sandberg, Anders, and Bostrom, Nick. 2011. *Machine Intelligence Survey*. Technical Report 2011-1. Future of Humanity Institute, University of Oxford.

Sandberg, Anders, and Savulescu, Julian. 2011. "The Social and Economic Impacts of Cognitive Enhancement." In *Enhancing Human Capacities*, edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane, 92-112. Malden, MA: Wiley-Blackwell.

Schaeffer, Jonathan. 1997. *One Jump Ahead: Challenging Human Supremacy in Checkers*. New York: Springer.

Schaeffer, J., Burch, N., Bjornsson, Y., Kishimoto, A., Muller, M., Lake, R., Lu, P., and Sutphen, S. 2007. "Checkers Is Solved." *Science* 317 (5844): 1518-22.

Schalk, Gerwin. 2008. "Brain-Computer Symbiosis." *Journal of Neural Engineering* 5 (1): P1-P15.

Schelling, Thomas C. 1980. *The Strategy of Conflict*. 2nd ed. Cambridge, MA: Harvard University Press.

Schultz, T. R. 2000. "In Search of Ant Ancestors." *Proceedings of the National Academy of Sciences of the United States of America* 97 (26): 14028-9.

Schultz, W., Dayan, P., and Montague, P. R. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593-9.

Schwartz, Jacob T. 1987. "Limits of Artificial Intelligence." In *Encyclopedia of Artificial Intelligence*, edited by Stuart C. Shapiro and David Eckroth, 1: 488-503. New York: Wiley. Schwitzgebel, Eric. 2013. "If Materialism is True, the United States is Probably Conscious."

Working Paper, February 8.

Sen, Amartya, and Williams, Bernard, eds. 1982. *Utilitarianism and Beyond*. New York: Cambridge University Press.

Shanahan, Murray. 2010. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. New York: Oxford University Press.

Shannon, Robert V. 2012. "Advances in Auditory Prostheses." *Current Opinion in Neurology* 25 (1): 61-6.

Shapiro, Stuart C. 1992. "Artificial Intelligence." In *Encyclopedia of Artificial Intelligence*, 2nd ed., 1: 54-7. New York: Wiley.

Sheppard, Brian. 2002. "World-Championship-Caliber Scrabble." *Artificial Intelligence* 134 (1-2): 241-75.

Shoemaker, Sydney. 1969. "Time Without Change." *Journal of Philosophy* 66 (12): 363-81. Shulman, Carl. 2010a. *Omohundro's "Basic AI Drives" and Catastrophic Risks*. San Francisco, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2010b. *Whole Brain Emulation and the Evolution of Superorganisms*. San Francisco, CA: Machine Intelligence Research Institute.

Shulman, Carl. 2012. "Could We Use Untrustworthy Human Brain Emulations to Make Trustworthy Ones?" Paper presented at the AGI Impacts conference 2012.

Shulman, Carl, and Bostrom, Nick. 2012. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." *Journal of Consciousness Studies* 19 (7-8): 103-30. Shulman, Carl, and Bostrom, Nick. 2014. "Embryo Selection for Cognitive Enhancement: Curiosity or Game-Changer?" *Global Policy* 5 (1): 85-92.

Shulman, Carl, Jonsson, Henrik, and Tarleton, Nick. 2009. "Which Consequentialism? Machine Ethics and Moral Divergence." In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference*, October 1st-2nd, University of Tokyo, Japan. Proceedings, edited by Carson Reynolds and Alvaro Cassinelli, 23-25. AP-CAP 2009.

Sidgwick, Henry, and Jones, Emily Elizabeth Constance. 2010. *The Methods of Ethics*. Charleston, SC: Nabu Press.

Silver, Albert. 2006. "How Strong Is GNU Backgammon?" *Backgammon Galore!* September 16. Retrieved October 26, 2013. Available at http://www.bkgm.com/gnu/AllAboutGNU.html#how_strong_is_gnu.

Simeral, J. D., Kim, S. P., Black, M. J., Donoghue, J. P., and Hochberg, L. R. 2011. "Neural Control of Cursor Trajectory and Click by a Human with Tetraplegia 1000 Days after Implant of an Intracortical Microelectrode Array." *Journal of Neural Engineering* 8 (2): 025027.

Simester, Duncan, and Knez, Marc. 2002. "Direct and Indirect Bargaining Costs and the Scope of the Firm." *Journal of Business* 75 (2): 283-304.

Simon, Herbert Alexander. 1965. *The Shape of Automation for Men and Management*. New York: Harper & Row.

Sinhababu, Neil. 2009. "The Humean Theory of Motivation Reformulated and Defended." *Philosophical Review* 118 (4): 465-500.

Slagle, James R. 1963. "A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus." *Journal of the ACM* 10 (4): 507-20.

Smeding, H. M., Speelman, J. D., Koning-Haanstra, M., Schuurman, P. R., Nijssen, P., van Laar, T., and Schmand, B. 2006. "Neuropsychological Effects of Bilateral STN Stimulation in Parkinson Disease: A Controlled Study." *Neurology* 66 (12): 1830-6.

Smith, Michael. 1987. "The Humean Theory of Motivation." *Mind: A Quarterly Review of Philosophy* 96 (381): 36-61.

Smith, Michael, Lewis, David, and Johnston, Mark. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society* 63: 89-174.

Sparrow, Robert. 2013. "In Vitro Eugenics." *Journal of Medical Ethics*. doi:10.1136/medethics-2012-101200. Published online April 4, 2013. Available at <http://jme.bmj.com/content/early/2013/02/13/medethics-2012-101200.full>.

Stansberry, Matt, and Kudritzki, Julian. 2012. Uptime Institute 2012 Data Center Industry Survey. Uptime Institute.

Stapledon, Olaf. 1937. *Star Maker*. London: Methuen.

Steriade, M., Timofeev, I., Durmuller, N., and Grenier, F. 1998. "Dynamic Properties of Corticothalamic Neurons and Local Cortical Interneurons Generating Fast Rhythmic (30-40 Hz) Spike Bursts." *Journal of Neurophysiology* 79 (1): 483-90.

Stewart, P. W., Lonky, E., Reihman, J., Pagano, J., Gump, B. B., and Darvill, T. 2008. "The Relationship Between Prenatal PCB

Exposure and Intelligence (IQ) in 9-Year-Old Children.” *Environmental Health Perspectives* 116 (10): 1416-22.

Sun, W., Yu, H., Shen, Y., Banno, Y., Xiang, Z., and Zhang, Z. 2012. “Phylogeny and Evolutionary History of the Silkworm.” *Science China Life Sciences* 55 (6): 483-96. Sundet, J., Barlaug, D., and Torjussen, T. 2004. “The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Scores of Norwegian Conscripts During Half a Century.” *Intelligence* 32 (4): 349-62.

Sutton, Richard S., and Barto, Andrew G. 1998. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

Talukdar, D., Sudhir, K., and Ainslie, A. 2002. “Investigating New Product Diffusion Across Products and Countries.” *Marketing Science* 21 (1): 97-114.

Teasdale, Thomas W., and Owen, David R. 2008. “Secular Declines in Cognitive Test Scores: A Reversal of the Flynn Effect.” *Intelligence* 36 (2): 121-6.

Tegmark, Max, and Bostrom, Nick. 2005. “Is a Doomsday Catastrophe Likely?” *Nature* 438: 754.

Teitelman, Warren. 1966. “Pilot: A Step Towards Man-Computer Symbiosis.” PhD diss., Massachusetts Institute of Technology.

Temple, Robert K. G. 1986. *The Genius of China: 3000 Years of Science, Discovery, and Invention*. 1st ed. New York: Simon & Schuster.

Tesauro, Gerald. 1995. “Temporal Difference Learning and TD-Gammon.” *Communications of the ACM* 38 (3): 58-68.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton, NJ: Princeton University Press.

Tetlock, Philip E., and Belkin, Aaron. 1996. "Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives." In *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, edited by Philip E. Tetlock and Aaron Belkin, 1-38. Princeton, NJ: Princeton University Press.

Thompson, Adrian. 1997. "Artificial Evolution in the Physical World." In *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, edited by Takashi Gomi, 101-25. Er '97. Carp, ON: Applied AI Systems.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., et al. 2006. "Stanley: The Robot That Won the DARPA Grand Challenge." *Journal of Field Robotics* 23 (9): 661-92.

Trachtenberg, J. T., Chen, B. E., Knott, G. W., Feng, G., Sanes, J. R., Welker, E., and Svoboda, K. 2002. "Long-Term In Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex." *Nature* 420 (6917): 788-94.

Traub, Wesley A. 2012. "Terrestrial, Habitable-Zone Exoplanet Frequency from Kepler."

Astrophysical Journal 745 (1): 1-10.

Truman, James W., Taylor, Barbara J., and Awad, Timothy A. 1993. "Formation of the Adult Nervous System." In *The Development of Drosophila Melanogaster*, edited by Michael Bate and Alfonso Martinez Arias. Plainview, NY: Cold Spring Harbor Laboratory.

Tuomi, Ilkka. 2002. "The Lives and the Death of Moore's Law." *First Monday* 7 (11). Turing, A. M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433-60. Turkheimer, Eric, Haley, Andreana, Waldron, Mary, D'Onofrio, Brian, and Gottesman, Irving I. 2003. "Socioeconomic Status Modifies Heritability of IQ in Young Children." *Psychological Science* 14 (6): 623-8.

Uauy, Ricardo, and Dangour, Alan D. 2006. "Nutrition in Brain Development and Aging: Role of Essential Fatty Acids." *Supplement, Nutrition Reviews* 64 (5): S24-S33.

Ulam, Stanislaw M. 1958. "John von Neumann." *Bulletin of the American Mathematical Society* 64 (3): 1-49.

Uncertain Future, The. 2012. "Frequently Asked Questions" The Uncertain Future. Retrieved March 25, 2012. Available at <http://www.theuncertainfuture.com/faq.html>.

U.S. Congress, Office of Technology Assessment. 1995. U.S.-Russian Cooperation in Space ISS-618. Washington, DC: U.S. Government Printing Office, April.

Van Zanden, Jan Luiten. 2003. On Global Economic History: A Personal View on an Agenda for Future Research. International Institute for Social History, July 23.

Vardi, Moshe Y. 2012. "Artificial Intelligence: Past and Future." *Communications of the ACM* 55 (1): 5.

Vassar, Michael, and Freitas, Robert A., Jr. 2006. "Lifeboat Foundation Nanoshield." Lifeboat Foundation. Retrieved May 12, 2012. Available at <http://lifeboat.com/ex/nanoshield>. Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and*

Engineering in the Era of Cyberspace, 11-22. NASA Conference Publication 10129. NASA Lewis Research Center. Visscher, P. M., Hill, W. G., and Wray, N. R. 2008. "Heritability in the Genomics Era: Concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255-66.

Vollenweider, Franz, Gamma, Alex, Liechti, Matthias, and Huber, Theo. 1998. "Psychological and Cardiovascular Effects and Short-Term Sequelae of MDMA ('Ecstasy') in MDMA- Naive Healthy Volunteers." *Neuropsychopharmacology* 19 (4): 241-51.

Wade, Michael J. 1976. "Group Selections Among Laboratory Populations of *Tribolium*." *Proceedings of the National Academy of Sciences of the United States of America* 73 (12): 4604-7.

Wainwright, Martin J., and Jordan, Michael I. 2008. "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning* 1 (1-2): 1-305. Walker, Mark. 2002. "Prolegomena to Any Future Philosophy." *Journal of Evolution and Technology* 10 (1).

Walsh, Nick Paton. 2001. "Alter our DNA or robots will take over, warns Hawking." *The Observer*, September 1. <http://www.theguardian.com/uk/2001/sep/02/medicalscience.genetics>.

Warwick, Kevin. 2002. *I, Cyborg*. London: Century.

Wehner, M., Oliker, L., and Shalf, J. 2008. "Towards Ultra-High Resolution Models of Climate and Weather." *International Journal of High Performance Computing Applications* 22 (2): 149-65.

Weizenbaum, Joseph. 1966. "Eliza: A Computer Program for the Study of Natural Language Communication Between Man And

Machine.” *Communications of the ACM* 9 (1): 36-45. Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to*

Calculation. San Francisco, CA: W. H. Freeman.

Werbos, Paul John. 1994. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. New York: Wiley.

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. 1986. “The Structure of the Nervous System of the Nematode *Caenorhabditis Elegans*.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 314 (1165): 1-340.

Whitehead, Hal. 2003. *Sperm Whales: Social Evolution in the Ocean*. Chicago: University of Chicago Press.

Whitman, William B., Coleman, David C., and Wiebe, William J. 1998. “Prokaryotes: The Unseen Majority.” *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578-83.

Wiener, Norbert. 1960. “Some Moral and Technical Consequences of Automation.” *Science* 131 (3410): 1355-8.

Wikipedia. 2012a, s.v. “Computer Bridge.” Retrieved June 30, 2013. Available at http://en.wikipedia.org/wiki/Computer_bridge.

Wikipedia. 2012b, s.v. “Supercomputer.” Retrieved June 30, 2013. Available at <http://en.wikipedia.org/wiki/Supercomputer>.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton Science Library. Princeton, NJ: Princeton University Press.

Winograd, Terry. 1972. *Understanding Natural Language*. New York: Academic Press.

Wood, Nigel. 2007. *Chinese Glazes: Their Origins, Chemistry and Re-creation*. London: A.& C. Black. World Bank. 2008. *Global Economic Prospects: Technology Diffusion in the Developing World* 42097. Washington, DC.

World Robotics. 2011. Executive Summary of 1. World Robotics 2011 Industrial Robots; 2. World Robotics 2011 Service Robots. Retrieved June 30, 2012. Available at [http://www. bara.org.uk/pdf/2012/world-robotics/Executive_Summary_WR_2012.pdf](http://www.bara.org.uk/pdf/2012/world-robotics/Executive_Summary_WR_2012.pdf).

World Values Survey. 2008. WVS 2005-2008. Retrieved 29 October, 2013. Available at <http://www.wvsevsdb.com/wvs/WVSAnalyzeStudy.jsp>.

Wright, Robert. 2001. *Nonzero: The Logic of Human Destiny*. New York: Vintage.

Yaeger, Larry. 1994. "Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or PolyWorld: Life in a New Context." In *Proceedings of the Artificial Life III Conference*, edited by C. G. Langton, 263-98. Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. Machine Intelligence Research Institute, San Francisco, CA, June 15. Yudkowsky, Eliezer. 2002. "The AI-Box Experiment." Retrieved January 15, 2012. Available at [http:// yudkowsky.net/singularity/aibox](http://yudkowsky.net/singularity/aibox).

Yudkowsky, Eliezer. 2004. Coherent Extrapolated Volition. Machine Intelligence Research Institute, San Francisco, CA, May.

Yudkowsky, Eliezer. 2007. "Levels of Organization in General Intelligence." In Artificial General Intelligence, edited by Ben Goertzel and Cassio Pennachin, 389-501. Cognitive Technologies. Berlin: Springer.

Yudkowsky, Eliezer. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global

Risk." In Global Catastrophic Risks, edited by Nick Bostrom and Milan M. Cirkovic, 308-45. New York: Oxford University Press.

Yudkowsky, Eliezer. 2008b. "Sustained Strong Recursion." Less Wrong (blog), December 5.

Yudkowsky, Eliezer. 2010. Timeless Decision Theory. Machine Intelligence Research Institute, San Francisco, CA.

Yudkowsky, Eliezer. 2011. Complex Value Systems are Required to Realize Valuable Futures. San Francisco, CA: Machine Intelligence Research Institute.

Yudkowsky, Eliezer. 2013. Intelligence Explosion Microeconomics, Technical Report 2013-1. Berkeley, CA: Machine Intelligence Research Institute.

Zahavi, Amotz, and Zahavi, Avishag. 1997. The Handicap Principle: A Missing Piece of Darwin's Puzzle. Translated by N. Zahavi-Ely and M. P. Ely. New York: Oxford University Press.

Zalasiewicz, J., Williams, M., Smith, A., Barry, T. L., Coe, A. L., Bown, P. R., Brenchley, P., et al. 2008. "Are We Now Living in the Anthropocene?" GSA Today 18 (2): 4-8.

Zeira, Joseph. 2011. "Innovations, Patent Races and Endogenous Growth." *Journal of Economic Growth* 16 (2): 135-56.

Zuleta, Hernando. 2008. "An Empirical Note on Factor Shares." *Journal of International Trade and Economic Development* 17 (3): 379-90.

Giải thích một số thuật ngữ

AI cấp độ con người: Một AI có thể phù hợp với hiệu suất trí tuệ của một người điển hình trong tất cả các lĩnh vực thực tế quan trọng. (Thuật ngữ này có sự mơ hồ đáng kể.)

AI công cụ: Một AI không giống như một tác tử, nhưng giống như một phiên bản phần mềm hiện đại và linh hoạt hơn. Đặc biệt, nó không hướng đến mục tiêu.

AI đúng đắn về đạo đức (MR): Một AI tìm cách làm những gì đúng về mặt đạo đức.

AI hạt giống: Một AI có khả năng ban đầu khiêm tốn có thể tự khởi động thành một AI ấn tượng bằng cách cải thiện kiến trúc của chính nó.

AI thần đèn: Một AI thực hiện một lệnh cấp cao, sau đó chờ đợi một lệnh khác.

AI tiên tri: Một AI chỉ trả lời các câu hỏi.

AI toàn năng: Một AI hoạt động tự chủ trên thế giới để theo đuổi các mục tiêu rộng lớn và có thể rất dài hơi.

Bồi đắp giá trị liên kết: Một cách tiếp cận vấn đề tải truyền giá trị, trong đó một cơ chế được chỉ định sẽ khiến AI chấp nhận các giá trị cuối cùng mới theo thời gian và theo một cách nào đó phụ thuộc vào kinh nghiệm của nó. (Lấy cảm hứng từ cách con người xuất hiện để đạt được các giá trị.)

Các phương pháp kiểm soát năng lực: Các chiến lược đưa ra nhằm tránh các kết quả không mong muốn bằng cách giới hạn những gì AI có thể làm.

Các phương pháp tương thưởng: Chiến lược kiểm soát AI bao gồm việc thiết lập môi trường của AI, sao cho nó có lý do công cụ ngay cả khi mục tiêu tối thượng của nó không phù hợp với giá trị của con người.

Cách tiếp cận chọn lọc tiến hóa với vấn đề truyền tải giá trị: Cách tiếp cận vấn đề truyền tải giá trị tìm cách có được AI với các giá trị mong muốn thông qua quá trình chọn lọc lặp lại, tương tự như quá trình chọn lọc tiến hóa tạo ra con người.

Cân nhắc quan trọng: Một ý tưởng hoặc lập luận có thể tiết lộ rõ ràng sự cần thiết không chỉ của một số điều chỉnh tiến trình nhỏ trong nỗ lực thực tế của chúng ta mà còn là một sự thay đổi lớn về hướng hoặc ưu tiên, ví dụ bằng cách “đổi dấu” của mức độ đáng mong muốn của sự can thiệp quan trọng.

Cắt cánh: Sự chuyển đổi từ một điều kiện trong đó chỉ có trí tuệ máy cấp độ con người sang một điều kiện có siêu trí tuệ toàn diện. Nó thường được đặc trưng bởi tốc độ: một quá trình “cắt cánh chậm” phải mất hàng thập kỷ hoặc hàng thế kỷ, quá trình “cắt cánh trung bình” phải mất hàng tháng hoặc hàng năm, và một quá trình “cắt cánh nhanh” chỉ tính bằng ngày hoặc ít hơn.

Công suất tối ưu hóa: Định lượng nỗ lực thiết kế được điều chỉnh về chất lượng đang được áp dụng để cải thiện trí tuệ của hệ thống.

Đặc tả trực tiếp: Một cách tiếp cận vấn đề kiểm soát, trong đó các lập trình viên tìm ra giá trị của con người, sau đó viết mã cho AI chứa rõ ràng các giá trị hoặc quy tắc tương ứng.

Điều biến giả lập: Bắt đầu với các bản giả lập não bộ có các động lực bình thường của con người và sửa đổi các động lực của chúng bằng cách sử dụng các phương pháp kỹ thuật số tương tự như thuốc, hoặc các phương tiện khác.

Gia cường: Cách tiếp cận để có được một siêu trí tuệ với các động lực mong muốn, bao gồm việc khởi đầu với một hệ thống đã có động lực thích hợp (ví dụ: một con người) và làm cho nó thông minh hơn, thay vì cố gắng thiết kế một hệ thống mục tiêu từ đầu.

Giả định hoàn thiện công nghệ: Nếu các nỗ lực phát triển khoa học công nghệ không hoàn toàn chấm dứt, con người sẽ có được toàn bộ năng lực cơ bản quan trọng có thể có được thông qua một số công nghệ khả thi.

Giả lập hoàn chỉnh não bộ: Trí tuệ máy được tạo ra bằng cách sao chép cấu trúc tính toán của não người.

Giam hãm vị nhân: Một hiện tượng được giả thuyết, trong đó AI nghĩ rằng nó có thể bị giam trong một bản mô phỏng, và vì vậy cố gắng hành xử nhằm được thưởng bởi các bộ mô phỏng.

Gói phần thưởng: Các điều khoản trong các mục tiêu được trao cho một AI phân bổ phần thưởng hoặc ảnh hưởng bổ sung cho những người đã giúp hiện thực hóa AI.

Hiện thực hóa sai lệch: Một cách hiệu quả để thỏa mãn mục tiêu của AI nhưng vi phạm ý định của các lập trình viên, những người đã xác định tiêu chí mục tiêu (ví dụ: đạt được mục tiêu khiến con người mỉm cười bằng cách làm tê liệt cơ mặt người thành một nụ cười vĩnh viễn).

Học giá trị: Cách tiếp cận vấn đề truyền tải giá trị, trong đó AI học các giá trị mà con người muốn nó theo đuổi.

Kết quả đa cực: Một điều kiện sau khi chuyển sang kỷ nguyên máy thông minh trong đó có nhiều tổ chức siêu trí tuệ cạnh tranh.

Khớp nối công nghệ: Mối quan hệ thời gian dự đoán được giữa hai công nghệ, sao cho việc phát triển một trong các công nghệ có xu hướng mạnh mẽ dẫn đến sự phát triển của công nghệ kia, như là một phiên bản tiền thân cần thiết, một ứng dụng không thể phủ nhận hoặc một bước xảy ra kế tiếp. Ví dụ, giả lập hoàn chỉnh não bộ có thể khớp nối với AI hình thái thần kinh, bởi vì phiên bản công nghệ nguyên thủy hơn mà công nghệ giả lập đòi hỏi có thể đã cho phép tạo ra AI lấy cảm hứng từ não bộ (và việc áp dụng công nghệ này để thực hiện điều đó được tưởng thưởng đáng kể).

Khung tạo động lực: Cách tiếp cận vấn đề truyền tải giá trị, trong đó AI hạt giống ban đầu được đưa ra các mục tiêu đơn giản mà sau đó được thay thế bằng các mục tiêu phức tạp hơn (liên kết giá trị) một khi AI đã phát triển đủ tài nguyên đại diện.

Luận đề hội tụ công cụ: Chúng ta có thể xác định các “giá trị hội tụ công cụ” rất hữu ích cho việc đạt được một loạt các mục tiêu tối thượng khả thể trong một loạt các môi trường khả thể của các tác tử thông minh.

Luận đề trực giao: Trí tuệ và mục tiêu tối thượng là trực giao: về nguyên tắc, bất kỳ mức độ trí tuệ nào có thể được kết hợp với bất kỳ mục tiêu tối thượng nào. Ngưỡng bền vững của thể đơn nhất thông thái: Một tập hợp các năng lực vượt qua ngưỡng thông minh khi và chỉ khi một hệ thống kiên trì và hiểu biết về rủi ro diệt vong sở hữu tập hợp năng lực này có thể xâm chiếm và thiết kế lại một phần lớn của vũ trụ có thể tiếp cận khi không gặp phải bất kỳ sự kháng cự hay đối địch thông minh nào.

Nguyên tắc lợi ích chung: Siêu trí tuệ chỉ nên được phát triển vì lợi ích của toàn nhân loại và phục vụ cho các lý tưởng đạo đức phổ quát.

Nguyên tắc phát triển công nghệ thiên lệch: Làm chậm sự phát triển của các công nghệ nguy hiểm và có hại, đặc biệt là những công nghệ làm tăng mức độ rủi ro diệt vong; và đẩy nhanh sự phát triển của các công nghệ có lợi, đặc biệt là những công nghệ làm giảm các rủi ro diệt vong do tự nhiên hoặc bởi các công nghệ khác.

Nguyên tắc tôn trọng nhận thức: Một siêu trí tuệ trong tương lai chiếm một điểm thuận lợi vượt trội về mặt nhận thức: những gì nó tin tưởng (có lẽ, về hầu hết các chủ đề) sẽ có nhiều khả năng đúng hơn so với niềm tin của chúng ta. Do đó, chúng ta nên tuân theo quan điểm của siêu trí tuệ bất cứ khi nào có thể. Phần cứng cấp độ con người: Phần cứng phù hợp với khả năng xử lý thông tin của não người.

Phần mềm cấp độ con người: Phần mềm phù hợp với hiệu quả thuật toán của não người, để thực hiện các nhiệm vụ của não người.

Phương pháp đóng hộp: Một phương pháp kiểm soát bao gồm việc thiết kế môi trường của AI để giảm khả năng nó tương tác với thế giới bên ngoài; ví dụ: chạy AI trong một mô phỏng thực tế ảo bị cô lập cẩn thận chỉ với một kênh liên lạc hạn chế với người giám sát của nó.

Phương pháp học tăng cường cho vấn đề kiểm soát: Cách tiếp cận mà trong đó, AI học để tối đa hóa một số khái niệm về phần thưởng tích lũy (trong đó tín hiệu phần thưởng được chỉ định hoặc quản lý bởi con người theo cách nào đó nhằm tạo ra hành vi phù từ phía AI).

Quan điểm ảnh hưởng đến cá nhân: Quan điểm cho rằng một người nên hành động vì lợi ích tốt nhất của tất cả những người đã tồn tại hoặc những người sẽ tồn tại độc lập với lựa chọn của họ (xem Quan điểm phi cá nhân).

Quan điểm phi cá nhân: Quan điểm cho rằng một người nên hành động vì lợi ích tốt nhất của mọi người, bao gồm cả những người có thể tồn tại do lựa chọn của người khác (xem quan điểm ảnh hưởng đến cá nhân).

Quy chuẩn gián tiếp: Cách tiếp cận vấn đề chọn lọc giá trị, thay vì trực tiếp tự xác định giá trị AI cần quảng bá, chúng tôi chỉ định một tiêu chí hoặc phương pháp mà AI có thể tuân theo, sử dụng tài nguyên trí tuệ của riêng mình để khám phá nội dung cụ thể của một tiêu chuẩn quy định ngầm.

Rủi ro đơn đoạn: Một rủi ro đến từ việc thực hiện chuyển đổi. Ở đây số lượng rủi ro không phải là một hàm đơn giản của việc chuyển đổi mất bao lâu. Ví dụ, đi qua một bãi mìn sẽ không an toàn hơn khi được thực hiện nhanh hơn.

Rủi ro trạng thái: Một rủi ro xuất phát từ một trạng thái nhất định, như vậy số lượng rủi ro là một hàm trực tiếp của thời gian. Ví dụ, trạng thái không có công nghệ để bảo vệ khỏi các tác động của tiểu hành tinh mang rủi ro tỷ lệ thuận với khoảng thời gian chúng ta ở trong trạng thái đó.

Siêu trí tuệ: Bất kỳ trí tuệ nào vượt quá hiệu suất nhận thức của con người trong hầu hết các lĩnh vực quan tâm.

Siêu trí tuệ chất lượng: Một hệ thống có tốc độ ít nhất tương đương trí tuệ của con người và thông minh hơn về mặt chất lượng.

Siêu trí tuệ tập thể: Một hệ thống bao gồm một số lượng lớn các yếu tố nhỏ hơn sao cho hiệu suất tổng thể trên nhiều lĩnh vực tổng quát vượt xa bất kỳ hệ thống nhận thức hiện tại nào.

Siêu trí tuệ tốc độ: Một hệ thống có thể làm tất cả những gì trí tuệ con người có thể làm, nhưng nhanh hơn rất nhiều.

Sự bùng nổ trí tuệ: Một sự kiện được giả thuyết, trong đó AI cải thiện nhanh chóng từ mức độ “tương đối khiêm tốn” của thành một mức trí tuệ siêu phàm (trong một quá trình thường được tưởng tượng là liên quan đến việc tự cải thiện đệ quy).

Sức chống đối: Mức độ khó khăn trong việc cải thiện một hệ thống.

Tài nguyên vũ trụ của loài người: Kho chứa tài nguyên vật chất trong vũ trụ mà một nền văn minh hoàn thiện có nguồn gốc Trái đất (xuất phát từ Trái đất ngày nay) có thể tiếp cận.

Tăng cường nhận thức: Cải thiện khả năng trí tuệ của hệ thống.

Thể đơn nhất: Một trật tự thế giới mà trong đó, ở cấp độ ra quyết định cao nhất, chỉ có một tác tử nắm quyền (mặc dù trong nội bộ nó có thể chứa nhiều phe phái và lợi ích); đó là, một điều kiện mà trong đó, các vấn đề phối hợp toàn cầu quan trọng hầu hết được giải quyết. Các ví dụ có thể bao gồm một nền dân chủ thế giới, một nhà độc tài thế giới không bị khuất phục hoặc một AI siêu trí tuệ đủ mạnh để trấn áp bất kỳ đối thủ tiềm năng nào.

Tội ác tâm trí: Việc ngược đãi các quá trình tính toán có liên quan về mặt đạo đức (trong các bản mô phỏng, hoặc trong AI, hoặc được tạo ra trong chất nền máy vì lý do công cụ).

Tự cải thiện đệ quy: Quá trình AI (có thể là AI hạt giống) lặp đi lặp lại cải thiện trí tuệ của chính mình – sử dụng trí tuệ ngày càng tăng

của nó để áp dụng sức mạnh tối ưu hóa ngày càng mạnh mẽ vào nhiệm vụ tự nâng cao nhận thức.

Tự hạn định: Cách tiếp cận vấn đề kiểm soát trong đó hệ thống động lực của AI được thiết kế để đưa ra tham vọng rất khiêm tốn cho AI, hạn chế phạm vi của những điều nó muốn can thiệp vào.

Ước chế phát triển: Một phương pháp kiểm soát bao gồm việc giới hạn các khả năng nội tại của AI, chẳng hạn như bằng cách hạn chế quyền truy cập thông tin hoặc giới hạn kỹ thuật đối với các năng lực nhận thức của AI.

Ưu thế chiến lược quyết định: Sự vượt trội chiến lược (bằng công nghệ hoặc phương tiện khác) đủ để một tác tử thống trị thế giới hoàn toàn.

Vấn đề truyền tải giá trị: Vấn đề của việc khiến cho AI theo đuổi các giá trị mà chúng ta muốn nó theo đuổi như những mục tiêu tối thượng.

Vấn đề uỷ thác-đại diện thứ hai: Vấn đề mà con người (bên uỷ thác), với mong muốn thiết kế một AI siêu trí tuệ (bên đại diện) để đại diện cho mình. Còn gọi là vấn đề kiểm soát.

Vấn đề uỷ thác-đại diện thứ nhất: Vấn đề nổi tiếng mà trong đó một người (bên uỷ thác) phải đối mặt khi chỉ định một người khác (bên đại diện) hành động vì lợi ích của người uỷ thác (ví dụ, trong mối quan hệ giữa chủ và nhân viên).

Yếu tố gia tốc phát triển cấu trúc vĩ mô: Một đòn bẩy tương tượng (được sử dụng trong các thí nghiệm tương tượng) sẽ thay đổi tốc độ phát triển các đặc tính cấu trúc vĩ mô (như công nghệ và động lực địa chính trị), trong khi không thay đổi tốc độ mà các vấn đề của con người ở cấp độ vi mô diễn ra.

Danh mục thuật ngữ này được đưa vào để bạn đọc tiện theo dõi, mặc dù không đầy đủ (nhiều thuật ngữ và khái niệm khu biệt được giới thiệu trong sách không được liệt kê tại đây) và cũng không được bố trí tốt (phần thảo luận trong sách nên được ưu tiên phòng một số khái niệm có thể được hiểu không rõ ràng). (Tôi xin cảm ơn Stephanie Zolayvar và Katja Grace của AI Impacts, Viện nghiên cứu trí tuệ máy vì đã biên soạn một danh sách và cho phép tôi chỉnh sửa. Xem phiên bản gốc tại <http://aiimpacts.org/ai-risk-terminology/>.)

Chỉ mục

A AI công cụ 255-263 định nghĩa 255

AI hạt giống 64, 65, 75, 137, 148, 149, 161, 162, 167, 168, 185, 201, 203, 204, 206, 207, 242, 257, 311, 316, 320-322, 329- 332, 346, 349, 352, 353, 358, 362, 374, 376, 400, 401, 447, 462, 464, 479

AI hình thái thần kinh 93, 395- 397, 403, 407, 408, 448, 513, 516

AI kẹp giấy 21

AI mạnh 49

AI-OUM, xem khái niệm về tính tối ưu

AI-RL, xem khái niệm về tính tối ưu

AI thần đèn 251-255 định nghĩa 251

AI toàn năng 251-255

AI-VL, xem khái niệm về tính tối ưu

Alan Turing 26, 56, 89

Albert Einstein 152

Alfred N. Whitehead 28

Anders Sandberg 12, 13

Arthur Samuel 39

B

backgammon 40, 316

ba điều luật robot học 237

Benny Hill 182

Bertrand Russell 155, 237

biểu diễn tường minh 346
bộ chứng minh định lý 44
bộ giải phương trình 44
bồi đắp giá trị 318, 319, 320, 346
bộ ngắt mạch tự động 47
bức màn vô thức 255, 420, 485

C

Cách mạng Công nghiệp 22, 23, 121, 144, 270, 274
Cách mạng Nông nghiệp 22, 23, 25, 121, 144, 437
các phương pháp kiểm soát, 218- 245, 337-338, 389, 393
xem thêm kiểm soát năng lực và lựa chọn động lực
các phương pháp tưởng thưởng 224-229, 244
hội nhập xã hội 224-226, 265-266, 267, 337-338
token phần thưởng mã hóa 227 can thiệp dòng mầm 85, 86, 88,
124, 460
xem thêm chọn lọc phôi Carl Shulman 229, 302 cấy
ghép não, xem cyborg C. elegans 74, 448
CEV, xem ý chí ngoại suy kết hợp chạy đua công nghệ 145, 146,
385
checkers 39
chế độ sai lỗi 202, 205, 206, 210, 211, 213, 215, 217
CHINOOK 39
chi phí thương lượng 306
chọn lọc di truyền 77, 81, 83, 86, 87, 99, 115, 388, 397
xem thêm tiến hóa
chọn lọc phôi 80, 81, 87, 88, 124, 449, 451
chủ nghĩa khoái lạc 238, 351, 366 chuỗi hướng mục tiêu 190

computronium, 176, 177, 211, 213, 239, 324, 366
xem thêm công suất tính toán công nghệ nano 166, 170, 172, 174,
179, 297, 304, 385, 386, 398, 467, 469, 512, 514
công suất tính toán, 34, 58, 71, 74, 108, 110, 114, 126, 132-134, 136,
140, 176, 230, 261, 283, 334, 376, 404, 416
xem thêm computronium và mái hiên phần cứng Cờ vây 42
cờ vua 38, 40, 42, 51, 55, 101, 163, 230, 441, 442

D

Daniel Dewey 12, 13, 496

DART 45

David Chalmers 12, 445

dây bẫy 222, 233, 234, 235, 482

Deep Blue 40

Deep Fritz 55

Derek Parfit 12, 473

Donald Knuth 39

Douglas Lenat 40

Dự án bộ gen người 420, 467

Dự án hệ thống máy tính thế hệ thứ năm 31

Dự án Manhattan 149, 152, 154

Dự án Mùa hè Dartmouth 27 dung lượng bộ nhớ 30, 132, 456

Dutch book 192

di thừa hạ tầng 211, 212, 214, 215, 259, 313, 377, 478

đặc tả trực tiếp 236, 237, 239, 241, 486

điểm kỳ dị 21, 23, 24, 97, 139, 437, 438, 463

xem thêm sự bùng nổ trí tuệ 21, 23, 24, 97, 139, 437, 438, 463

điểm Schelling 249, 308, 483

điều biến giả lập 335-337, 347
điều khoản lợi nhuận bất ngờ 421, 422
điều kiện Malthus 274, 276, 277
định luật Moore 58, 60, 62, 134, 138, 140, 462, 486
xem thêm công suất tính toán 58, 60, 62, 134, 138, 140, 462
đóng hộp 221-224, 244, 265-266
cách ly thông tin 222
cách ly vật lý 221-222
động lực chạy đua, xem chạy đua công nghệ
đúng đắn về đạo đức 255, 363, 364, 365, 366, 506, 507, 508, 545
được phép về đạo đức 366, 508, 509
đường cơ sở con người 116, 117, 135, 137, 138, 140, 147, 463
đường cơ sở nền văn minh 117

E

Eliezer Yudkowsky 13, 162, 331, 353, 447, 478, 496
ELIZA 29
Enigma 154, 467
Eric Drexler 12, 13, 398
Eugene Wigner 152
Eurisko 40, 442

F

Flash Crash 47, 48
Franklin D. Roosevelt 152
FreeCell 41
Freeman Dyson 176

G

gánh nặng đột biến 82

Garry Kasparov 12, 40
Gerard K. O'Neill 176
Gerry Tesauro 40
gia cường 236, 241, 242, 24
Giả định hoàn thiện công nghệ 383 giải ô chữ 41
giải thưởng Field 423, 424, 457
giả lập hoàn chỉnh não bộ 54, 63, 66, 67, 69, 71-75, 98, 99, 102, 114, 115, 125-127, 140, 148, 150, 151, 177, 187, 288, 319, 333, 335, 337, 394-397, 400, 403-407, 419, 447, 448, 455, 492, 512-516, 547
giao dịch thuật toán 46, 47, 48, 443
giao diện não-máy tính, 89-95, 100, 148, 242
xem thêm cyborg
giao diện người-máy, xem cyborg giao thức trao đổi khóa Diffie-Hellman 145
giả thuyết mô phỏng 230, 231, 471
H
hàm thích nghi 33, 59, 60, 446
xem thêm tiến hóa
hàm thoả dụng 36, 156, 157, 174, 190, 204, 214, 228, 311-313, 327-329, 334, 347, 415, 416, 493, 494, 497, 499, 500, 501, 519, 520
Hannah Arendt 182
Hans Berliner 40
Hans Moravec 56
Harry Truman 152
Hãy làm điều tôi muốn 367, 509
hedonium 239, 366, 367
hệ thống hỗ trợ quyết định, xem thêm AI công cụ

hệ thống kiểm soát kho hàng 45

hiệu suất giảm dần 78, 122, 123, 157, 196, 460, 519

HLMI, xem trí tuệ máy, cấp độ con người

học giá trị 321, 323, 325, 326, 327, 329, 330, 334, 347, 497, 499, 500

học máy 33, 34, 37, 39, 42, 44, 45, 48, 63, 65, 113, 137, 151, 162, 167,
209, 258, 289, 316, 337, 432, 443, 462, 466

học tăng cường 40, 63, 209, 316, 317, 326, 337, 346, 396, 432, 469,
478, 494, 495, 548

hồi hải mã 93, 94

I

I. J. Good 441

Internet 45, 91, 97, 125, 131, 134, 140, 151, 166, 169, 172, 173, 223,
234, 247, 402, 456

Isaac Asimov 237

Isaac Newton 107

IVF, xem chọn lọc phôi

J

Jeopardy! 41, 130

Johannes Kepler 43

John McCarthy 28, 38, 49

John Rawls 485

John Stuart Mill 351

John von Neumann 437, 468

Jonathan Schaeffer 39

K

kết nối luận 33

khái niệm về tính tối ưu 37, 311,

326-329

học giá trị (AI-VL) 329

học tăng cường (AI-RL) 326

khuếch đại độ thỏa dụng quan sát (AI-OUM) 327

tác tử Bayes 35-37

khai thác dữ liệu 44, 46, 388, 515

khủng hoảng bản thể học 248, 330

khung tạo động lực 320, 321, 323

kịch bản đa cực 225, 267, 278

kiểm soát năng lực 218, 221, 226, 229, 243, 247, 248, 263, 266, 310,
311, 320, 338, 480, 502, 545

kinh tế thế giới 22, 23, 24, 25, 118, 136, 148, 274, 437, 462, 469, 484

L

lần chuyển đổi thứ hai 297, 298, 404, 405, 408

lập thời gian biểu 45

Leó Szilárd 152

lễ thường 25, 42, 238, 442

Liên Hợp Quốc 155, 156, 158, 419, 420

liệu pháp gen sinh dưỡng 85, 88

Logic Theorist 28

Logistello 40

lồng Faraday 222

Louise Brown 87

lựa chọn động lực 221, 225, 227, 236, 241-243, 247, 248, 250, 263,
281, 305, 313, 321, 335, 338, 371, 480, 481, 493, 502 định nghĩa 221

luận đề hội tụ công cụ 187, 188, 198, 199

luận đề trực giao 186, 187, 188, 198, 199, 472, 473

lý thuyết automata 27

lý thuyết chọn lọc quan sát, xem vị nhân lý thuyết quyết định 349, 370, 373, 374, 376, 379, 380, 474, 493, 509, 510

xem thêm khái niệm về tính tối ưu

lý thuyết trò chơi 155, 268

M

mã hoá RSA 145

mái hiên phần cứng 134, 401, 404, 492, 516, 518

mạng lưới thần kinh 27, 32, 33, 34, 63, 67, 93, 126, 235, 396, 433, 440, 461, 462

Mark Oliphant 152

Marvin Minsky 49, 438, 478

Matt Ginsberg 41 máy trẻ con 56, 64

MegaEarth 107

Mikhail Gorbachev 154

mô hình điện toán thần kinh 67, 74, 75

xem thêm giả lập hoàn chỉnh

não bộ (WBE) và AI hình thái thần kinh

mô hình Hodgkin-Huxley 60 mô hình logic, xem Trí tuệ Nhân tạo Thời xưa (GOFAL)

mô hình Markov ẩn 43

mục tiêu dựa trên lý do 368 mục tiêu tự hạn chế 211

N

neuron McCulloch-Pitts 396 ngã rẽ xảo trá 204, 205, 220 ngôn ngữ hình thức 31, 247

ngôn ngữ tự nhiên 42, 130, 238, 246, 365, 433

ngữ nghĩa hiệu lực nhân quả 331

ngữ nghĩa ngoại tham chiếu 331
ngưỡng bền vững của thể đơn nhất thông thái 178, 179, 180, 472
Nguyên tắc lợi ích chung 421, 422
Nguyên tắc phát triển công nghệ thiên lệch 382, 384, 547
Nguyên tắc tôn trọng nhận thức 352, 547
nhân bản vô tính 84
nhận dạng giọng nói 43, 45, 433 nhận dạng khuôn mặt 44, 433
nhận dạng ký tự 44
nhận thức luận 36, 165, 349, 370, 374-376, 379, 380, 428, 498, 499, 506, 509, 510, 511
nhận thức sinh học 54, 89, 99, 100, 242, 388, 429
những lập luận phán đoán 399
Nicolaus Copernicus 43
Nils Nilsson 49-51
nội dung mục tiêu 190, 248, 346, 371, 372, 374, 380, 510
nỗ lực thiết kế, xem công suất tính toán

P

Patrick Winston 49
Paul Christiano 12, 334, 347, 434
Peter Norvig 49, 443
phát tín hiệu xã hội 192, 193, 495
phê chuẩn 266, 370, 377, 378, 379, 479
phương pháp Hail Mary 347, 501
poker 41
Proverb 41

Q

qualia, xem ý thức

quan điểm ảnh hưởng đến cá nhân

408, 409, 516, 548

quan điểm phi cá nhân 382, 386, 390, 400, 430

quét, xem giả lập hoàn chỉnh não bộ (WBE)

quy chuẩn gián tiếp 236, 241, 254, 349, 352, 353, 354, 363, 369, 372,
438, 510

R

Ray Kurzweil 453

Robin Hanson 12, 23, 268, 456

robot 29, 42, 43, 45, 67, 71, 95, 166, 169, 170, 202, 203, 222, 237, 268,
279, 396, 435, 456, 466, 482, 494

Ronald Reagand 154

rủi ro diệt vong 178-180, 259, 303, 384-386, 388, 390, 391, 394, 400-
402, 405, 406, 408, 409, 424, 425, 427, 429, 431, 432, 477, 512, 514, 516-
518, 547

rủi ro đơn đoạn 390

rủi ro trạng thái 390

S

sai lỗi ác tính 206, 211, 217

Scrabble 41

Sebastian Thrun 49

Shakey 29

SHRDLU 29

siêu năng lực 102, 161, 164-170, 172-174, 180, 181, 205, 228, 250,
469, 472, 506

các loại 165-166

siêu sinh vật 299, 300, 492, 493

siêu trí tuệ chất lượng 102, 108, 109, 111

định nghĩa 108

siêu trí tuệ tập thể 80, 95, 97, 102, 105-108, 111, 148, 164

định nghĩa 104

siêu trí tuệ tốc độ 102, 103, 104, 111, 139, 174, 455, 456, 458

định nghĩa 102

sự bùng nổ tổ hợp 29, 30, 34, 35, 37, 94, 260

sự bùng nổ trí tuệ 24, 26, 27, 65, 116, 119, 137, 139, 141, 150-153, 168, 198, 218, 220, 229, 257, 271, 276, 332, 343, 379, 386, 387, 392-394, 400-402, 405, 408, 409, 413, 414, 418, 419, 422, 424, 425, 430, 433, 463, 467, 477, 482, 487, 513, 516, 517, 519

sức chống đối 116, 121, 122, 123, 124, 125, 127, 128, 129, 131, 132, 134, 135, 136, 137, 138, 139, 140, 162, 401, 462

sự thoái hoá không đáng kể 32

sự tuyệt chủng của con người, xem

rủi ro diệt vong

T

Tác tử Bayes, 35-37, 212, 222

xem thêm tác tử nhân tạo và khái niệm về tính tối ưu

tác tử nhân tạo, xem thêm tác tử Bayes

tải lên, xem giả lập hoàn chỉnh não bộ (WBE)

tài nguyên vũ trụ 177, 179, 181, 193, 196-198, 230, 350, 358, 362, 379, 416, 431, 479, 505 tầm tiếp cận gián tiếp 110, 162, 178

tầm tiếp cận trực tiếp 111, 162

tăng cường nhận thức 85, 89, 93, 100, 123, 162, 165, 191-193, 242, 324, 344, 388, 389, 391-394, 397, 406, 429, 449, 450, 451, 477,

tăng cường sinh học, 76-89, 100, 242-243, 388
xem thêm tăng cường nhận thức tăng trưởng đạo đức 357
tăng trưởng kinh tế 273, 274, 437, 463, 512
tàu thăm dò von Neumann 174, 175, 194, 195, 196
TD-Gammon 40
TextRunner 130, 131, 343, 346, 356, 377, 508
thảm họa giả thuyết Rieman 211
tổng hợp ADN 172 240
Tribolium castaneum 263
Thành viên Taliban 359
trí tuệ con người 26, 38, 58, 59,
thất nghiệp 120, 270, 273, 487
thể đơn nhất 142, 147, 155, 156, 158-160, 178-181, 194, 196, 199,
204, 267, 268, 297-299, 302, 306, 309, 385, 402, 406, 418, 465, 466, 471,
472, 475, 476, 489, 512, 515, 519, 520, 547
thiên kiến quy nạp 34, 35
thiết kế thể chế 338, 345
thị giác máy 37, 73
thuật toán di truyền 32, 33, 34, 41, 57, 58, 62, 396, 400
thuật toán lan truyền ngược 32
thực tại ảo 67, 103, 195, 230, 279, 286, 287
thụ đắc tài nguyên 195, 196, 197, 200, 324, 476
thuốc thông minh 77, 123, 449
thụ tinh trong ống nghiệm, xem chọn lọc phôi
thuyết cướp đoạt theo kiểu Pascal 373, 510
thuyết đặt cược của Pascal 373
tìm kiếm tự nghiệm 30

tìm kiếm vết cạn 29, 30

tín hiệu phần thưởng 209, 210, 211, 326, 346, 548

tội ác tâm trí 215, 259, 315, 330, 60, 92, 109, 110, 162, 172, 234, 268, 288, 389, 403, 424, 549

trí tuệ máy cách mạng, xem sự bùng nổ trí tuệ cấp độ con người (HLMI) 25-27, 51-52, 66-75, 135-136

xem thêm trí tuệ nhân tạo trí tuệ nhân tạo

các nhà tiên phong 25-27, 48-49

chạy đua vũ trang 119, 410-412

dự báo quá mức về 25

lịch sử của 27-47

tương lai của 48-53

Trí tuệ Nhân tạo Thời xưa 32, 56

truyền tải giá trị 310, 314, 315, 317, 319, 320, 322, 323, 328, 330, 331, 334, 335, 345, 346, 347, 348, 499, 545, 547

tự cải thiện đệ quy 65, 138, 168, 242, 429 549

xem thêm AI hạt giống tự động hóa 46, 48, 73, 82, 169, 173, 202, 227, 270, 434, 439

tự hạn định 236, 240, 241, 247, 255, 313, 321, 346, 371

tự kỷ 109, 252

U

ức chế phát triển 229-233, 244

ưu sinh 76, 77, 85, 86, 451, 471

ưu thế chiến lược quyết định 156, 158, 167, 181, 193, 198-201, 205, 206, 208, 209-211, 214, 215, 217, 221, 225, 231, 236, 250, 252, 265-267, 297, 298, 319, 350, 358, 376, 418

Y

ý chí ngoại suy kết hợp 334, 353-355, 358, 361, 363, 364, 366, 371, 505, 510

định nghĩa 353

yếu tố gia tốc phát triển cấu trúc vĩ mô 391, 393

yếu tố gia tốc phổ quát 389

V

vấn đề AI toàn diện 93, 131, 164, 246

vấn đề ủy thác-đại diện 218-221

Vernor Vinge 23, 97, 455

vị nhân 218, 230, 244, 264, 266, 370, 480, 546

vị thế đạo đức 216, 280, 291, 337, 343, 451, 491, 505

W

Watson 41, 130, 442

WBE 403, 404, 405, 406, 407, 408, 409, 515, 516

wireheading 209-211, 228, 316, 326, 327, 346

X

xác định kiểu gen 77

xử lý biểu tượng thứ cấp, xem

kết nối luận xấp thuật toán 290