Mục lục

Lời tựa

Cuốn sách này viết về những gì loài người cần học hỏi để có thể phát triển mạnh mẽ trong Kỉ nguyên Trí tuệ nhân tạo. Luận điểm bao trùm của tôi là: chúng ta cần học cả những kĩ năng và kiến thức có thể bổ trợ cho những gì AI làm tốt, lẫn những kĩ năng và kiến thức cho phép chúng ta tương tác với AI một cách hiệu quả, năng suất và có trách nhiệm.

Tôi không thể đề cập thấu đáo mọi thứ chúng ta cần học, nhưng có thể giới thiệu sơ lược cho người đọc về các chủ đề khác nhau. Điều này đủ để mọi người quyết định xem có nên tìm hiểu thêm thông qua các nguồn tài liệu tôi trích dẫn hay không – bao gồm nhiều video hướng dẫn miễn phí trên YouTube, như những video được dẫn nguồn ở cuối mỗi chương. Cuốn sách này có thể giúp độc giả xây dựng một chương trình tự học toàn diện, góp phần quan trọng vào việc chuẩn bị cho những gì sắp tới. Để sử dụng các câu lệnh gợi ý (prompt) và bảng biểu dễ dàng hơn, độc giả có thể sao chép và dán chúng tại địa chỉ: www.routledge.com/9781032686653.

Những cuốn sách trước đây của tôi về các chủ đề ứng dụng trải dài từ cách thiết kế các bảng hiển thị trực quan, cách thiết kế các bài thuyết trình bằng bài trình chiếu (slideshow), cách thiết kế các bài tập học tập chủ động cho các khóa học trực tuyến, cho đến

cách thiết kế, giảng dạy và đánh giá học tập chủ động với sự hỗ trợ của AI. Trong mọi trường hợp, tôi đều tập trung vào phương pháp và kĩ thuật, chứ không phải nội dung. Cuốn sách này thì khác: Nó tập trung vào những kĩ năng và kiến thức cụ thể mà tất cả chúng ta cần học để phát triển rực rỡ trong Kỉ nguyên AI.

Tôi chủ ý dùng từ "phát triển mạnh mẽ". Mục tiêu của tôi ở đây không chỉ là cung cấp một cuốn cẩm nang sinh tồn, cũng không phải là đưa ra những gợi ý hữu ích hay các phương pháp tốt nhất. Thay vào đó, tôi đã phát triển một phương pháp tiếp cận có hệ thống và mang tính kiến tạo: phương pháp này sẽ cho phép bạn làm việc với AI để hoàn thành các mục tiêu của riêng mình. Tôi mong muốn giúp độc giả có một cuộc sống tốt đẹp hơn nhờ sử dụng AI.

Với chủ đề và những khuyến nghị trong cuốn sách này, sẽ là đạo đức giả nếu từ chối sử dụng AI để hỗ trợ mình trong quá trình viết. Và trên thực tế, tôi đã thường xuyên vận dụng "Vòng lặp khuếch đại nhận thức" được mô tả trong Chương 2. Tôi nhận thấy AI đặc biệt hữu ích trong việc gợi ý các ví dụ và phép so sánh, biên tập nội dung, cũng như cung cấp cái nhìn tổng quan về các chủ đề và tên tuổi của những nhân vật đóng góp quan trọng. Tôi không dùng AI để lập dàn ý hay lập cấu trúc cho bất kì phần nào của cuốn sách, và tôi cũng chưa bao giờ cho phép nó đưa ra quyết định cuối cùng: Trong mọi trường hợp, tôi đều biên tập và chỉnh sửa lại những gì AI gợi ý – tôi đã thực hành đúng những gì mình chia sẻ trong Chương 2. Ngoại lệ duy nhất cho điều này là khi tôi cung cấp các bản ghi chép về các tương tác với AI, đó là những bản ghi nguyên văn về những gì đã thực sự diễn ra.

Dù vậy, những đóng góp của AI cho cuốn sách này nhỏ bé hơn nhiều so với sự đóng góp của rất nhiều bạn bè, đồng nghiệp và thành viên gia đình: Như mọi khi, vợ tôi, Tiến sĩ Robin Rosenberg, đã đưa ra những lời khuyên sâu sắc, những phê bình đầy thấu đáo và các gợi ý biên tập chi tiết. Các con tôi là Justin, David và

Neil đều đã dành thời gian tranh luận với tôi và chỉnh sửa cho tôi những điểm kĩ thuật – cả ba đều có công việc đòi hỏi sự am hiểu sâu sắc về AI, và tôi đã nhận được sự hỗ trợ từ chuyên môn của các con theo nhiều cách mà có lẽ chính chúng không nhận ra hết. Tôi cũng xin cảm ơn Tiến sĩ Anne Marie Ward và Heyu Huang, những người đã đưa ra những nhận xét sâu sắc về các bản thảo đầu tiên, góp phần định hình cho dự án cuối cùng. Tôi cũng cần phải một lần nữa cảm ơn Tiến sĩ Beth Callaghan, người bạn và đồng nghiệp của tôi tại Active Learning Sciences, người đã đọc bản thảo và đóng vai trò như "người phản biện" một cách hiệu quả. Những phê bình mang tính xây dựng và sáng tạo của cô ấy đã khiến cuốn sách này trở nên tốt hơn rất nhiều. Và Tiến sĩ Kacey Warren đã cung cấp một bản phê bình vô cùng sắc sảo và chi tiết cho một phiên bản trước đó, giúp tôi tập trung hơn và diễn đạt rõ ràng hơn. Cô ấy đã mang đến góc nhìn sắc bén của một triết gia, khiến tôi phải suy nghĩ kĩ càng về những điều mình đang trình bày – và cùng những điều khác, chính điều đó đã khiến tôi quyết định loại bỏ hẳn một chương. Tôi cũng phải nói rõ rằng, mọi lỗi lầm hay sai sót còn lại hoàn toàn là trách nhiệm của tôi. Tôi xin cảm ơn Niki Gallagher Garcia vì đã hỗ trợ nghiên cứu và hiệu đính. Và, tất nhiên, tôi cũng cảm ơn Tori Sharpe và Alison Macfarlane tại Routledge cùng biên tập viên Tom Bedford, những người đã luôn kiên nhẫn và hỗ trợ tôi ở mọi bước đường.

CHƯƠNG 1:

Con người trong thời đại AI

OpenAI, một công ty khởi nghiệp tại Thung lũng Silicon, đã phát hành phần mềm mang tính đột phá ChatGPT vào ngày 30 tháng 11 năm 2022. Loại hình trí tuệ nhân tạo (AI) mới này ngay lập tức gây ra một hỗn hợp cảm xúc gồm sốc, kinh ngạc, phấn khích, và – tôi nghĩ có thể nói là – cả nỗi sợ hãi và sự e ngại. Chỉ trong vòng hai tháng, hơn 100 triệu người đã trở thành người dùng tích cực, đạt tốc độ phổ cập công nghệ mới nhanh kỉ lục.[1] Mọi người đã nhiệt tình đón nhận loại AI mới này vì lí do chính đáng: nó có khả năng viết rất tốt về nhiều chủ đề khác nhau, và các thế hệ sau của nó – từ các công ty như Anthropic, Google, Meta và OpenAI – có thể làm được mọi thứ, từ việc tạo ra một giáo trình hợp lí cho một khóa học quản trị kinh doanh, đến việc soạn một bài giảng đạo có thể chấp nhận được, gợi ý những tiêu đề phim hấp dẫn dựa trên vài gạch đầu dòng về cốt truyện, phân tích bản mô tả công việc để suy luận về các năng lực cần thiết, sáng tác thơ, tổng hợp ý tưởng, cho đến viết những bài luận có thể được xuất bản về nhiều chủ đề khác nhau. Hơn nữa, những AI này tiếp thu chỉ dẫn rất tốt: Nếu không thích bản nháp ban đầu, ta có thể yêu cầu chúng cải thiện kết quả đầu ra. Ví dụ, ta có thể yêu cầu chúng điều chỉnh văn phong trang trọng hơn, cung cấp thêm ví dụ,

tập trung hơn vào một khía cạnh cụ thể, và/hoặc thay đổi mức độ phức tạp. Ngay cả phiên bản ChatGPT đầu tiên cũng đã làm tất cả những điều này cực kì tốt.[2]

Nhìn chung, các AI mới này giỏi phân tích, đánh giá, tổng hợp và sáng tạo ra những điều mới dựa trên kiến thức đã có. Thực tế, chúng ta có thể cung cấp cho chúng một bảng tiêu chí chấm điểm và yêu cầu chúng sử dụng bảng tiêu chí đó để chấm các bài luận của người học, và kết quả chấm điểm mà chúng đưa ra khá tương đồng với cách một giảng viên con người chấm điểm.[3] Các AI mới cũng có thể diễn giải và tạo ra hình ảnh. Chúng ta có thể tải lên một bức ảnh chụp cánh cửa tủ lạnh đang mở của mình và yêu cầu AI gợi ý các công thức nấu ăn dựa trên những gì đang có, và nó sẽ đáp ứng ngay. Chúng ta cũng có thể yêu cầu nó hiển thị hình ảnh về các giống mèo khác nhau, và lập tức những hình ảnh đó sẽ xuất hiện. Thật vậy, chúng ta thậm chí có thể yêu cầu một hình ảnh về một con vật không có thật, nửa mèo nửa chó, và nó dễ dàng tạo ra con vật lai đó. Hầu như tuần nào cũng có công bố về một khả năng mới đáng kinh ngạc nào đó mà AI có thể làm được, từ phát hiện ung thư và dự đoán chứng mất trí nhớ, cho đến việc có thể chuyển đổi các đoạn văn mô tả thành infographic ấn tượng hay các video sống động như thật.[4]

Các AI mới này là ví dụ về trí tuệ nhân tạo sinh dữ liệu (generative artificial intelligence – loại AI có khả năng tạo ra nội dung mới dựa trên dữ liệu khổng lồ mà chúng được huấn luyện). Trong cuốn sách này, chúng ta tập trung vào các AI sinh dữ liệu như vậy, và từ đây trở đi tôi sẽ gọi đơn giản là "AI". Sự tiến bộ trong lĩnh vực này đã diễn ra với tốc độ đáng kinh ngạc kể từ khi ChatGPT ra mắt. Các công ty AI giới thiệu những đổi mới nhanh đến mức nhiều người cảm thấy không kịp theo dõi. Thật vậy, chưa đầy sáu tháng sau khi ChatGPT được phát hành, hơn 1.000 nhà nghiên cứu về AI và những người làm việc trong các lĩnh vực liên quan đã kí một bản kiến nghị yêu cầu tạm dừng mọi nghiên cứu

về các mô hình AI lớn hơn, có năng lực hơn trong sáu tháng. Mục đích là để cho các nhà nghiên cứu – và xã hội nói chung – có cơ hội suy ngẫm về cách chúng ta sẽ kiểm soát một công nghệ như vậy.[5] Mọi người đã lo ngại về đủ mọi chuyện, từ việc AI sẽ khiến con người mất việc cho đến viễn cảnh chúng chuyển sang chế độ "Kẻ Hủy Diệt" toàn diện, thâu tóm thế giới và tiêu diệt loài người. Với sự cạnh tranh khốc liệt giữa các công ty AI mới tham gia thị trường, bản kiến nghị này chắc chắn đã thất bại ngay từ đầu, nhưng nó phản ánh rõ ràng mức độ đột phá và khả năng làm rung chuyển thế giới của các AI mới – cũng như mức độ lo lắng của nhiều người về chúng. Tiếp sau đó, nhiều chính phủ hiện đã ban hành các quy định quản lí việc phát triển và sử dụng AI,[6] và chắc chắn sẽ có thêm nhiều quy định như vậy nữa trong tương lai.

Trong bối cảnh tiến bộ nhanh chóng như vậy, rủi ro khi viết sách về AI là nó sẽ sớm trở nên lạc hậu. Đó là lí do tại sao tôi không tập trung vào bản thân các AI, mà thay vào đó là vào người dùng – con người. Cuốn sách này viết về những gì con người cần học để có thể phát triển rực rỡ trong một thế giới vận hành cùng AI.

Bản chất của "quái thú"

Để phát triển rực rỡ trong một thế giới vận hành cùng AI, chúng ta cần có kiến thức cơ bản về cách thức hoạt động của các AI sinh dữ liệu. Kiến thức đó có thể cho phép chúng ta phát huy điểm mạnh và hạn chế những điểm yếu của chúng. Vì vậy, hãy bắt đầu với một cái nhìn tổng quan nhanh về các đặc điểm cơ bản của những AI mới này.

Nhiều đặc điểm của AI bắt nguồn từ chính thiết kế của các hệ thống phần mềm. Để thấy rõ điều này, chúng ta hãy bắt đầu bằng cách so sánh các AI mới với những chương trình máy tính truyền thống mà tất cả chúng ta đều biết, chẳng hạn như Adobe Acrobat, Apple Mail, Microsoft Excel và Microsoft Word. Hãy tưởng tượng

một tờ giấy và một cây bút chì. Tờ giấy lưu trữ các dạng biểu đạt, chẳng hạn như từ ngữ và hình vẽ; còn cây bút chì tạo ra và chỉnh sửa những biểu đạt đó theo những quy tắc nhất định (ví dụ: ngữ pháp đối với văn bản và luật phối cảnh đối với hình ảnh). Các chương trình máy tính truyền thống cũng tương tự: Chúng lưu trữ dữ liệu ở những vị trí cụ thể trong bộ nhớ máy tính, sau đó thao tác trên dữ liệu đó theo các quy tắc cụ thể. Ví dụ: một chương trình bảng tính tiêu chuẩn (như Microsoft Excel hoặc Google Sheets) hoạt động trên các chữ số được lưu trữ và có thể áp dụng cùng một bộ quy tắc, chẳng hạn như phép cộng và phép nhân, cho bất kì tập hợp chữ số nào. Nếu nó lưu trữ các số 20212 và 1640, nó có thể cộng hoặc nhân chúng, và điều tương tự cũng được áp dụng với 175 và 11, hay bất kì số nào khác. Các chữ số cụ thể là gì không quan trọng; miễn là chúng là chữ số, thì cỗ máy có thể áp dụng các quy tắc của số học lên chúng. Và chương trình áp dụng các quy tắc này một cách đáng tin cậy với độ chính xác cao, luôn cho cùng một kết quả với cùng một phép toán trên cùng các chữ số, bất kể nó thực hiện bao nhiêu lần.

Ngược lại, các AI mới là một loại hệ thống máy tính khác, không có cơ sở dữ liệu riêng biệt chứa các biểu diễn dạng kí hiệu (như chữ số và từ ngữ) để rồi tạo ra và xử lí chúng theo các quy tắc được lưu trữ sẵn (như các quy tắc số học và ngữ pháp). Thay vào đó, những AI này dựa trên mạng nơ-ron nhân tạo (artificial neural network – một mô hình tính toán phức tạp lấy cảm hứng từ cách các nơ-ron liên kết trong bộ não người). Các mạng nhân tạo này bao gồm một tập hợp các đơn vị gọi là "nơ-ron nhân tạo", thường lên đến hàng triệu đơn vị. Các nơ-ron nhân tạo này được sắp xếp thành các lớp, và mỗi đơn vị được kết nối với nhiều đơn vị khác ở các lớp liền kề. Các kết nối này có thể mang tính kích thích (dương) hoặc ức chế (âm). Nếu một nơ-ron có kết nối kích thích đến một nơ-ron khác, kích hoạt nơ-ron đầu tiên sẽ dẫn đến việc nó cố gắng kích hoạt nơ-ron thứ hai. Ngược lại, nếu một nơ-ron có kết nối ức chế với một nơ-ron khác, việc kích hoạt nơ-ron đầu tiên

sẽ dẫn đến việc nó cố gắng ngăn cản nơ-ron thứ hai bị kích hoạt. Có rất, rất nhiều kết nối như vậy giữa các nơ-ron nhân tạo. Hơn nữa, các kết nối này khác nhau về cường độ: cường độ càng lớn thì một nơ-ron sẽ kích hoạt hoặc ức chế nơ-ron kia càng mạnh mẽ, tùy thuộc vào kết nối là kích thích hay ức chế.

Trong mạng nơ-ron, bất kì một mẩu thông tin cụ thể nào cũng tương ứng với một kiểu mẫu (pattern) nhất định về cường độ của các kết nối giữa nhiều đơn vị nơ-ron. Thông tin không được định vị cục bộ như trong chương trình máy tính truyền thống. Thay vào đó, thông tin được phân tán trên nhiều kết nối. Thật vậy, bất kì một kết nối nào cũng có thể là một phần của nhiều kiểu mẫu khác nhau lưu trữ những mẩu thông tin khác nhau – hơi giống như cách một chữ cái có thể là một phần của nhiều từ trong trò chơi ô chữ. Hơn nữa, mạng nơ-ron không hoạt động theo cách của các chương trình máy tính tiêu chuẩn. Toàn bộ mạng nơ-ron hoạt động như một kiểu bộ lọc khổng lồ: Một tập hợp "các đơn vị đầu vào" được kích thích từ bên ngoài (ví dụ, như khi chúng ta đặt câu hỏi cho AI), điều này sẽ kích hoạt hoặc ức chế các đơn vị được kết nối, tùy theo bản chất của các kết nối. Làn sóng kích hoạt lan truyền về phía trước, đôi khi dội ngược lại, len lỏi qua mạng lưới và cuối cùng tạo ra một kết quả đầu ra cụ thể. Kết quả đầu ra cụ thể từ một đầu vào cụ thể phụ thuộc vào bản chất và cường độ của các kết nối giữa các đơn vị.

Điều gì quyết định cường độ của các kết nối? Nói ngắn gọn là: việc huấn luyện (training). Khi một AI được huấn luyện để xử lí ngôn ngữ, ban đầu mạng được cung cấp một lượng lớn văn bản, và nó cố gắng dự đoán các "token" kế tiếp. Token là đơn vị văn bản nhỏ nhất mà mạng xử lí (chẳng hạn như các từ ngắn, một phần của từ, hoặc kí hiệu). Mô hình dựa vào các token đứng trước để dự đoán token tiếp theo. Việc dự đoán này được thực hiện tự động, và vì mạng được cung cấp toàn bộ các từ và câu, nó có thể kiểm tra xem mình đoán đúng hay không và tự điều chỉnh cường độ giữa

các kết nối liên quan để sửa lỗi. Một khi mạng đã trở nên rất giỏi trong việc dự đoán như vậy, chúng sẽ được huấn luyện trên các tài liệu cụ thể hơn để "tinh chỉnh" (fine tune). Sau đó, con người tham gia vào vòng huấn luyện cuối cùng. Các huấn luyện viên con người này thực hiện quy trình "Học tăng cường từ phản hồi của con người" (Reinforcement Learning from Human Feedback - RLHF) để khuyến khích các kết quả đầu ra nhất định và không khuyến khích những kết quả khác. Nếu các huấn luyện viên con người hài lòng với một câu trả lời, họ sẽ khiến mạng lưới tăng cường các kết nối đã dẫn đến câu trả lời đó; ngược lại, nếu không hài lòng, họ sẽ làm suy yếu các kết nối tích cực và tăng cường các kết nối tiêu cực. Đây là nguồn gốc của các "rào chắn" (guardrail – những cơ chế được thiết lập để ngăn chặn mạng cung cấp thông tin hoặc lời khuyên nguy hiểm, chẳng hạn như giúp người dùng chế tạo bom hoặc phạm tội).

Các AI mới này thường có một "kiến trúc GPT". Chữ "G" là viết tắt của "Generative" (Sinh dữ liệu); mô hình không chỉ đơn giản tra cứu thông tin đã lưu trữ trước đó, như cách Google hay chức năng "Tìm kiếm" (Find) trong một chương trình xử lí văn bản hoạt động, mà thay vào đó, nó dựa vào các kiểu mẫu kết nối để tạo ra các phản hồi mới lạ. Chữ "P" là viết tắt của "Pre-trained" (Được huấn luyện trước); mô hình đã được huấn luyện từ trước, thường là dựa trên gần như mọi thứ có trên Internet, và thường được bổ sung thêm các nguồn khác. Và chữ "T" là viết tắt của "Transformer" (Bộ biến đổi); đối với ngôn ngữ, mạng sẽ diễn giải các token và chú ý đến vị trí xuất hiện cùng nhau của hai (hay nhiều) token trong câu, biến đổi đầu vào thành các đặc tả (specification) về ý nghĩa của các token cụ thể và mối quan hệ của chúng trong câu. Quá trình biến đổi này cho phép mạng tạo ra ảo giác về việc "hiểu được" ý nghĩa của đầu vào.

Đây là một cái nhìn tổng quan rất nhanh về một chủ đề phức tạp, nhưng hi vọng là đủ cho các mục đích hiện tại của chúng ta.

Nhiều nhà nghiên cứu có kiến thức sâu rộng đã cung cấp các bài hướng dẫn chi tiết về mạng nơ-ron và kiến trúc GPT ([7,8,9]).

Mặc dù các hệ thống AI sinh dữ liệu có khả năng tạo ra và diễn giải hình ảnh dựa trên các nguyên tắc cơ bản tương tự như các Mô hình Ngôn ngữ Lớn (Large Language Models - LLM) mà chúng ta đã thảo luận cho đến nay, chúng tôi tập trung vào LLM vì những AI này có nhiều khả năng nhất sẽ thay đổi một cách cơ bản cách chúng ta sống và làm việc.

Sống và làm việc trong thế giới vận hành cùng AI

Chúng ta đã có thể sử dụng Internet để truy cập phần lớn – nếu không muốn nói là hầu hết – kho tàng tri thức của nhân loại, được tích lũy qua suốt chiều dài lịch sử. Đối với hầu hết các mục đích, "đám mây" (cloud) có thể đóng vai trò như một kho lưu trữ kí ức mở rộng cho bộ não chúng ta; nó lưu giữ thông tin để chúng ta không cần phải tự mình làm điều đó. Hơn nữa, AI hiện đã giỏi ít nhất là ngang bằng với hầu hết con người ở nhiều kĩ năng nhận thức đòi hỏi xử lí thông tin, chẳng hạn như tư duy phản biện, tư duy sáng tạo, giải quyết các vấn đề được xác định rõ ràng, đưa ra các quyết định có cấu trúc chặt chẽ, nhận dạng các quy luật/mô hình, và đưa ra nhiều kiểu dự đoán. AI đã có thể thực hiện nhiều tác vụ nhận thức, và chúng sẽ ngày càng trở nên tốt hơn khi các mô hình mới được phát hành. AI có thể xử lí rất nhiều thông tin, vì vậy chúng ta sẽ không cần phải tự mình làm điều đó.

Những quan sát này dẫn đến những câu hỏi sâu sắc: Nếu chúng ta có thể tìm thấy bất kì thông tin nào trên đám mây khi cần, và AI có thể xử lí thông tin đó theo ý muốn của chúng ta, thì tại sao chúng ta phải bận tâm lĩnh hội những kiến thức và kĩ năng đó? Thay vào đó, chúng ta nên học gì? Chúng ta cần biết gì để phát triển rực rỡ trong thế giới mới đang hình thành này? Đây là một quan sát đơn giản đi thẳng vào trọng tâm của những vấn đề này:

Con người giỏi ứng phó trong các tình huống mở đòi hỏi phải xem xét đến bối cảnh (xem thêm[10,11,12]). Trong các tình huống mở, chúng ta không biết trước yếu tố nào sẽ đóng vai trò trung tâm, định hình tình huống đó. Trên thực tế, các yếu tố mới có thể bất ngờ xuất hiện, chẳng hạn như chuông báo cháy kêu hay một cuộc biểu tình trên đường phố, có thể biến một chuyến đi thông thường đến bưu điện thành một trải nghiệm đầy thử thách. Và "bối cảnh" đề cập đến những yếu tố định hình hoặc bao quanh một tình huống nhất định, ảnh hưởng đến cách chúng ta diễn giải tình huống đó. Một tiếng động lớn vang lên sẽ mang lại cảm nhận khác nhau vào giữa đêm khuya tĩnh mịch so với giữa một buổi chiều rực rỡ ánh sáng, và chúng ta có thể bật cười nếu ai đó lỡ lời trong một cuộc trò chuyện thân mật nhưng sẽ cảm thấy ngượng nghịu nếu họ nói điều tương tự trong một bài điếu văn. Con người có thể thích ứng nhanh chóng khi tình huống thay đổi theo những cách không lường trước được, và chúng ta cực kì giỏi trong việc điều chỉnh những gì mình nhận thức, suy nghĩ, cảm nhận và hành xử tùy thuộc vào bối cảnh. Hơn nữa, chúng ta có thể học hỏi để trở nên giỏi hơn nữa ở những kĩ năng này (ví dụ[13]).

Mặc dù AI rất hiệu quả trong các tình huống được xác định rõ ràng, chẳng hạn như đưa ra lời khuyên về việc diễn giải điểm số của các bài kiểm tra cụ thể, chúng lại gặp khó khăn khi tình huống mang tính mở và đòi hỏi phải xét đến bối cảnh.[14, 15, 16] Theo lời của Giáo sư Đại học Harvard Gary King (trao đổi cá nhân), AI giỏi nội suy (ước tính các giá trị *bên trong* phạm vi dữ liệu đã biết), nhưng kém ngoại suy (dự đoán các giá trị *bên ngoài* phạm vi dữ liệu đã biết).

Những khó khăn mà AI gặp phải với các tình huống như vậy không đơn thuần là kết quả của việc chúng không được cập nhật đủ thường xuyên. Thật ra, cách con người diễn giải các tình huống và sự kiện không chỉ dựa trên những trải nghiệm quá khứ, mà còn phụ thuộc vào cách bộ não và cơ thể con người vận hành. Thực tế,

chúng ta dựa vào những khía cạnh đặc thù của não bộ và cơ thể mình, điều đó có thể tạo ra những linh cảm và trực giác – những điều mà AI cần phải ngoại suy vượt ra ngoài tập dữ liệu mà nó đã được huấn luyện.

Để làm rõ điều này một cách cụ thể, chúng ta hãy xem xét các nghiên cứu kinh điển của nhóm nghiên cứu Antonio Damasio.[17, 18, 19, 20, 21, 22, 23] Họ đã thiết kế một trò chơi sử dụng bốn bộ bài, và mỗi lá bài cho biết mức tiền lời hoặc lỗ. Mục tiêu là tối đa hóa lợi nhuận bằng cách tìm ra những bộ bài nào có khả năng mang lại lời nhiều hơn so với lỗ. Các lá bài ở hai bộ A và B mang lại khoản lời cao ngay lập tức nhưng lại có những lá dẫn đến thua lỗ rất lớn, khiến tổng kết sau cùng là thua lỗ. Ngược lại, các lá bài ở hai bộ C và D tuy mang lại lợi nhuận nhỏ hơn nhưng lại ít thua lỗ hơn, dẫn đến kết quả chung là có lời theo thời gian.

Sau vài lượt rút bài, người tham gia thường thích rút từ bộ A và B vì lợi nhuận ban đầu hấp dẫn, nhưng rồi họ dần tích lũy thua lỗ đều đặn. Sau khoảng 20 đến 40 lượt, họ bắt đầu chuyển sang rút ở bộ C và D. Điều thú vị nhất là những gì diễn ra khi họ đã chơi sâu vào trò chơi và chuẩn bị rút từ bộ A hoặc B – những bộ có khả năng gây lỗ: Lúc này, cơ thể họ xuất hiện các phản ứng sinh lí như lòng bàn tay đổ mồ hôi, một dấu hiệu của sự lo lắng. Có vẻ như người chơi đã có một "linh cảm" rằng có điều gì đó không ổn với bộ A và B, và họ cảm nhận được điều này từ trước khi ý thức hoàn toàn về những gì đang xảy ra.

Những linh cảm như vậy dường như phát sinh từ một phần cụ thể của não bộ, đó là vỏ não trước trán bụng giữa, nằm ở phần dưới, chính giữa, phía trước của não bộ. Thật vậy, những bệnh nhân bị tổn thương cấu trúc não này không bao giờ thể hiện các phản ứng cảm xúc mang tính dự đoán trước và không bao giờ học được cách chuyển sang các bộ bài tốt hơn. Dựa trên những phát hiện ở các bệnh nhân bị tổn thương não, các nhà nghiên cứu đã suy luận rằng vỏ não trước trán bụng giữa đóng một vai trò

then chốt trong quá trình ra quyết định vô thức này. Cấu trúc não này dường như có chức năng liên kết chéo giữa các phản ứng cảm xúc và kết quả của quá trình suy luận logic, từ đó tạo ra những linh cảm và trực giác thúc đẩy chúng ta hành động theo những cách nhất định từ rất sớm, trước cả khi chúng ta có thể lí giải vì sao mình lại đưa ra những quyết định đó.

Vì không có cảm xúc hay cơ thể nên AI không thể sử dụng loại cơ chế này. Những linh cảm và trực giác này có thể xuất hiện trong các tình huống mở, nơi bối cảnh là quan trọng – chẳng hạn như trong các trò chơi bài hư cấu mới lạ – và do đó, AI không thể được huấn luyện trước về tất cả các tình huống đó.

Hơn nữa, việc chúng không có cơ thể, hormone và những yếu tố tương tự cũng đồng nghĩa rằng các AI hiện nay còn tồn tại một loạt những giới hạn khác. Hãy xem xét một nhận định sâu sắc của triết gia người Viên Ludwig Wittgenstein,[24] người đã nhận xét rằng: "Ngay cả khi một con sư tử có thể nói, chúng ta cũng không thể hiểu được nó." Điều này nảy sinh bởi vì sư tử có một "hình thái sống" khác với con người. Ví dụ, chúng chạy bằng bốn chân thay vì đứng thẳng bằng hai chân, có móng vuốt dài và sắc thay cho ngón tay, di chuyển giữa lùm cỏ cao, vân vân. Những khác biệt này khiến chúng có xu hướng "phân chia" thế giới về mặt khái niệm theo cách khác với con người, và do đó, ngay cả khi chúng có từ ngữ, thì các khái niệm nền tảng bên dưới cũng sẽ không tương ứng với các khái niệm của con người.

Chúng ta có thể đưa ra lập luận tương tự đối với AI: Chúng không có cơ thể, không có hormone, không trải nghiệm sự mệt mỏi, vân vân. Do đó, chúng không thể hiểu đầy đủ trải nghiệm làm người là như thế nào. Chúng có thể mô phỏng kiến thức đó dựa trên các tập dữ liệu huấn luyện của mình, và nội suy trong phạm vi tập dữ liệu đó, nhưng không thể dễ dàng ngoại suy sang các tình huống mở đòi hỏi phải xem xét bối cảnh. Hơn nữa, mặc dù tất cả con người đều là thành viên của cùng một loài, chúng ta không

giống hệt nhau, và những khác biệt này khiến mỗi người chúng ta nhìn nhận sự việc ít nhất là hơi khác một chút so với người khác. Vì vậy, AI sẽ không "phân chia" thế giới giống hệt như bất kì một con người cụ thể nào, và chúng sẽ gặp khó khăn khi cố gắng hiểu từng người trong các tình huống mở đòi hỏi phải tính đến bối cảnh. Tôi đã yêu cầu GPT-4 cung cấp các ví dụ về tình huống mở nơi bối cảnh đóng vai trò quan trọng. Đây là những gì nó đã tạo ra:

- ***Đàm phán thỏa thuận và hợp đồng.*** Trong đàm phán, khả năng nắm bắt nhanh chóng không khí trong phòng, hiểu phản ứng của đối phương và điều chỉnh chiến lược của chúng ta cho phù hợp có thể dẫn đến kết quả thuận lợi hơn. Bối cảnh tức thời như ngôn ngữ cơ thể, giọng điệu và cách chọn từ có thể cung cấp những thông tin quan trọng.
- ***Đưa ra phản hồi cho đồng nghiệp.*** Khi đưa ra phản hồi, điều quan trọng là phải nhạy bén với tâm trạng của đồng nghiệp, hoàn cảnh xung quanh và phản ứng của họ khi cuộc trò chuyện diễn ra.
- ***Quản lí không khí lớp học với tư cách là giáo viên.*** Giáo viên phải điều chỉnh phương pháp giảng dạy và biện pháp can thiệp dựa trên sự tham gia, mức độ hiểu bài và hành vi của học sinh.
- ***Biểu diễn sân khấu hoặc âm nhạc.*** Nghệ sĩ biểu diễn phải nhạy bén với phản ứng của khán giả và không khí chung. Những điều chỉnh nhanh chóng, chẳng hạn như thay đổi danh sách tiết mục hoặc tương tác với khán giả, có thể nâng cao đáng kể chất lượng buổi biểu diễn.
- ***Tiến hành các hoạt động ngoại giao.*** Các nhà ngoại giao phải nhận thức sâu sắc về các sắc thái văn hóa, bầu không khí chính trị và các tín hiệu tinh tế trong các cuộc trò chuyện.
- ***Cung cấp dịch vụ khách hàng.*** Chuyên viên dịch vụ khách hàng phải điều chỉnh cách tiếp cận của họ dựa trên phản ứng và cảm xúc tức thời của khách hàng.

- ***Xử lí các tình huống khẩn cấp và khủng hoảng.*** Trong các tình huống khẩn cấp, chẳng hạn như thiên tai, tai nạn hoặc các tình huống y tế nguy kịch, các quyết định nhanh chóng, phù hợp với bối cảnh có thể quyết định sự sống chết.
- ***Tiến hành các hoạt động chiến đấu và quân sự.*** Binh lính và chỉ huy phải liên tục nhận biết tức thời về môi trường xung quanh và thích ứng nhanh chóng với các điều kiện thay đổi trên chiến trường.
- ***Tham gia các hoạt động thể thao.*** Vận động viên và huấn luyện viên phải nhận thức sâu sắc về bối cảnh tức thời trong quá trình thi đấu, bao gồm hành động của đối thủ, điều kiện môi trường và tình trạng thể chất của chính họ.
- ***Thực hiện phẫu thuật.*** Bác sĩ phẫu thuật cần phản ứng rất nhanh nhạy với bối cảnh tức thời trong một ca mổ. Những thay đổi trong tình trạng của bệnh nhân, các biến chứng không lường trước và các phản ứng khác nhau đối với việc điều trị đòi hỏi việc ra quyết định nhanh chóng và mang tính thích ứng.
- ***Thực thi pháp luật.*** Nhân viên thực thi pháp luật khi ứng phó với các vụ việc phải kịp thời đánh giá và thích ứng với các tình huống thay đổi nhanh chóng, cân bằng giữa nhu cầu hành động tức thời với việc cân nhắc về an toàn và tính hợp pháp.
- ***Đánh giá các bản tin.*** Một yếu tố then chốt trong việc xác định một bản tin có chính xác hay không là đánh giá xem sự kiện được đưa tin có phù hợp với bối cảnh xung quanh hay không.

Các AI hoạt động độc lập không có khả năng thực hiện tốt các loại nhiệm vụ này. Do đó, việc chúng ta học hỏi các kĩ năng và kiến thức cần thiết để thực hiện chúng là hợp lí, bao gồm cả việc học cách tận dụng AI để giúp chúng ta làm điều đó. Tuy nhiên, cần lưu ý rằng những hạn chế về nhận thức và cảm xúc của con người – bao gồm cả các thiên kiến – có thể cản trở chúng ta thực hiện tốt bất kì nhiệm vụ nào đã nêu ở trên. Vì vậy, ta không chỉ cần học các kĩ năng và kiến thức cụ thể, mà còn cần nhận biết và quản lí những hạn chế của bản thân, như sẽ được thảo luận trong phần sau cuốn sách này.