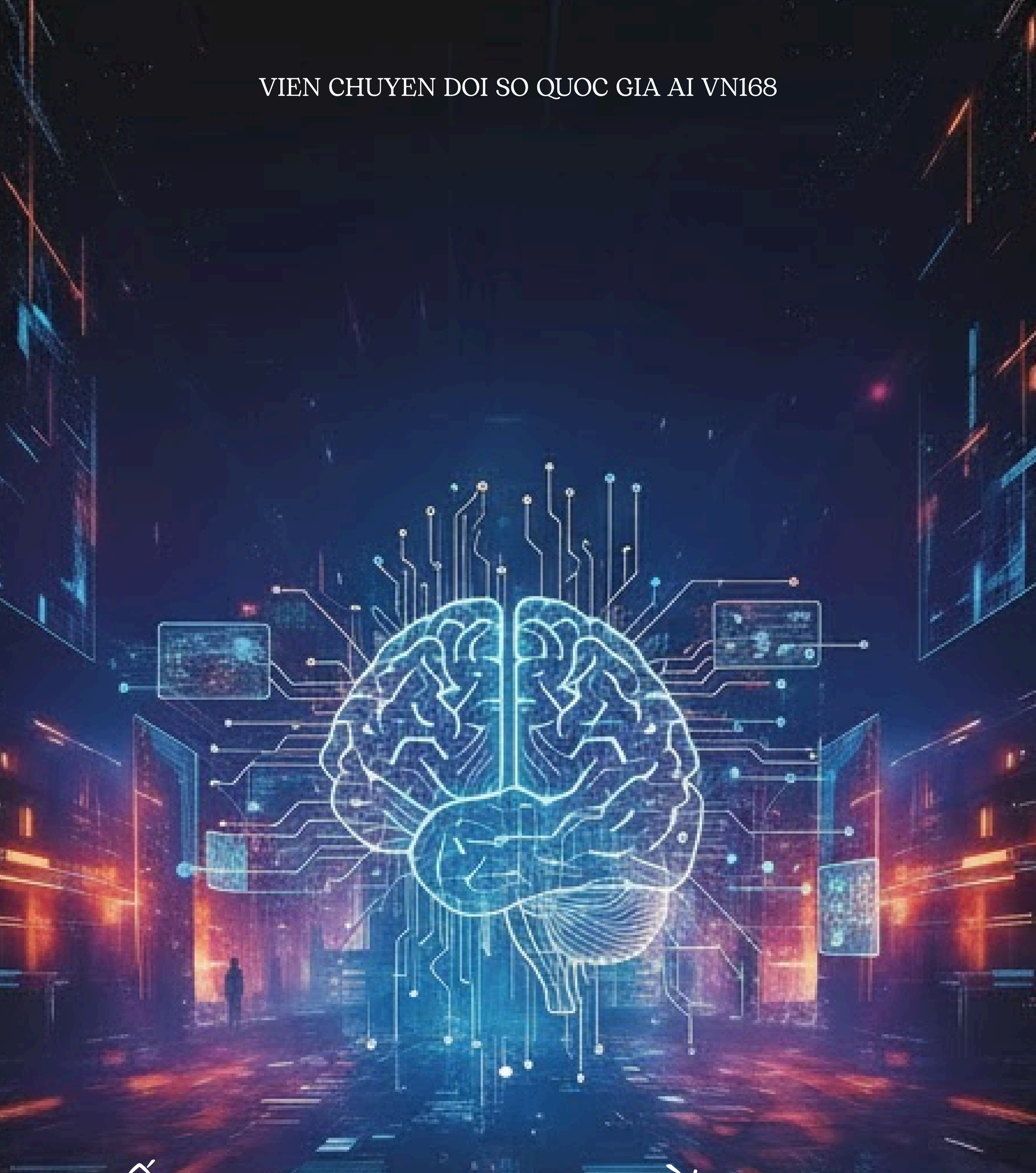
VIEN CHUYEN DOI SO QUOC GIA AI VN168



SỐNG TRONG THỜI ĐẠI AI

21 bài học để thích nghi

ThS. Bùi Thị Hải

MỤC LỤC

LỜI MỞ ĐẦU

Cuốn sách này đến với tôi từ những trải nghiệm rất đời thường khi làm việc cùng cơ quan, doanh nghiệp và cộng đồng: AI xuất hiện ngày một nhiều, đôi khi chỉ như một nút bấm "gợi ý", một hệ thống "tự động", một báo cáo "phân tích nhanh" - nhỏ thôi, nhưng rồi dần dần nó được tin tưởng và đưa vào những quyết định quan trọng. Tôi nhận ra điều khiến mình lo không phải vì AI quá cao siêu, mà vì nó đi vào cuộc sống quá êm: mọi thứ trơn tru đến mức ta dễ quên phải hỏi "vì sao", "đúng chưa", "ai chịu trách nhiệm nếu sai". Vì thế, tôi chọn chắp bút không để ca ngợi AI, cũng không để cảnh báo cực đoan, mà để trao cho bạn một chiếc la bàn: biết đặt câu hỏi đúng trước khi tin, biết nhìn thấy rủi ro trước khi trả giá, biết phân biệt năng lực thật với ảo tưởng trơn tru, và biết giữ phần trách nhiệm con người ở đúng vị trí của nó.

AI không bước vào đời sống của chúng ta như một phát minh ồn ào làm đổi thay mọi thứ chỉ sau một đêm. Nó đến âm thầm, len qua những khe nhỏ của thói quen: một gợi ý trong ô tìm kiếm, một đoạn tin nhắn được tự động hoàn chỉnh, một cuộc gọi được phân loại, một khoản vay được chấm điểm, một hồ sơ được ưu tiên xử lý. Khi mọi thứ vận hành trơn tru, ta gọi đó là tiện lợi và gần như quên mất đằng sau sự tiện lợi ấy có một cơ chế đang âm thầm định hình lựa chọn của mình. Chỉ đến khi hệ thống trượt khỏi đường ray - khi nó đề xuất sai, "đoán nhầm" con người, hoặc đẩy một quyết định sang phía bất lợi - ta mới nhận ra điều đáng bận tâm nhất không nằm ở việc máy móc giỏi đến đâu, mà ở chỗ năng lực ấy có thể được nhân bản với tốc độ công nghiệp: một quyết định được tự động hóa có thể được lặp lại hàng triệu lần; một sai lầm nhỏ có thể lan rộng thành một sai lầm hệ thống; một định kiến mơ hồ có thể trở thành chuẩn mực vận hành; và khi chi phí thử - sai lầm giảm xuống, tốc độ lan truyền tăng lên. xã hội bước vào một vùng rủi ro mới - nơi hậu quả không đến từ một cú ngã lớn, mà đến từ vô số bước trượt nhỏ, liên tục, khó nhìn thấy.Vì thế, hiểu AI không phải để hoảng loạn, cũng không phải để sùng bái; hiểu AI là để giữ mình tỉnh táo trước một công nghệ vừa hữu ích, vừa cso sức khuếch đại, và đôi khi nguy hiểm theo những cách rất "bình thường".

Muốn nói chuyện bình tĩnh về AI, ta cần một nền tảng đủ chắc để không bị mê hoặc bởi bề mặt trơn tru của công cụ. Nền tảng ấy bắt đầu từ câu hỏi tưởng như giản dị nhưng luôn gây tranh cãi: trí tuệ là gì, và ta đang dùng tiêu chuẩn nào để gọi một hành vi là "thông minh". Từ đó, ta cần nhìn thấy vì sao lịch sử AI rẽ thành hai truyền thống lớn: một truyền thống cổ điển tin vào luật lệ, biểu tượng và suy luận có cấu trúc; và một truyền thống hiện đại tin vào dữ liệu, học từ mẫu hình và sức mạnh của quy mô. Hai con đường này không chỉ khác kỹ thuật; chúng khác cả cách hình dung về trí tuệ, khác cách kỳ vọng vào máy, và vì thế khác cả kiểu rủi ro mà chúng tạo ra. Kế tiếp là những câu hỏi triết học vẫn treo lơ lửng như chiếc bóng đi cùng mọi tiến bộ: một hệ thống trả lời trôi chảy có nghĩa là nó "hiểu" không, hay chỉ đang thao tác trên dấu hiệu; một hệ thống hành động đúng trong nhiều tình huống có đồng nghĩa với "ý thức" không, hay chỉ là tối ưu hóa rất giỏi; và khi máy tham gia vào đời sống xã hội, ai chịu trách nhiệm cho sai sót - người thiết kế, người triển khai, hay những "quy tắc vô hình" của hệ thống? Cuối cùng, nền tảng ấy sẽ không trọn vẹn nếu ta bỏ qua robot - tức trí tuệ gắn với cơ thể và cảm biến - bởi khi AI bước ra khỏi màn hình để chạm vào thế giới vật lý, mọi thứ trở nên thật hơn, khó hơn, và cũng nghiêm khắc hơn: dữ liệu không còn sạch, môi trường không còn ổn định, sai số không còn chỉ là con số, và "thông minh" không còn là nói đúng, mà là hành động đúng và an toàn trong một thế giới đầy nhiễu.

Cuốn sách này bắt đầu từ những điều nền tảng ấy không phải để làm AI trở nên bí hiểm, mà để gỡ bỏ lớp sương mù quanh nó: giúp bạn nhìn AI như một hệ thống quyết định, một cỗ máy khuếch đại, và một tấm gương phản chiếu cách chúng ta định nghĩa tri thức, quyền lực và trách nhiệm - để từ đó, ta có thể sử dụng công nghệ bằng sự hiểu biết, thay vì để công nghệ âm thầm sử dụng mình.

1. **Trí tuệ: một từ quen nhưng lại khó hơn ta nghĩ**

Trí tuệ là một từ quá quen thuộc đến mức ta thường dùng nó như một nhãn dán: ai học nhanh, làm việc giỏi, nói năng sắc sảo, giải quyết vấn đề gọn gàng - ta gọi người đó "thông minh". Nhưng khi thử tháo nhãn ấy ra để nhìn vào bên trong, ta sẽ gặp một sự thật lạ: trí tuệ không phải một vật thể có thể cầm nắm, cũng không phải một công thức có thể chốt lại bằng một câu định nghĩa. Trí tuệ giống như một thước đo vừa khoa học vừa xã hội, vừa khách quan vừa mang màu sắc giá trị. Nó thay đổi theo những gì một cộng đồng coi trọng, theo những kỹ năng một thời đại cần để sống sót và thăng tiến, theo bối cảnh văn hóa và lịch sử. Bởi vậy, hỏi mười người "trí tuệ là gì" và nhận về mười câu trả lời không phải vì con người mơ hồ, mà vì mỗi người đang vô thức đứng trên một chiếc thang khác nhau để đo. Có người đo trí tuệ bằng tri thức: biết nhiều là thông minh. Có người đo bằng tốc độ: phản ứng nhanh là thông minh. Có người đo bằng khả năng suy luận: lập luận chặt chẽ là thông minh. Có người đo bằng khả năng sống: xoay xở, thích nghi, không gục ngã trước biến cố là thông minh. Thậm chí có người đặt trí tuệ vào vùng của cảm xúc và đạo đức: biết thấu cảm, biết kiểm soát bản thân, biết sống tử tế cũng là một dạng thông minh - một dạng thông minh mà xã hội thường nhận ra muộn, nhưng lại phụ thuộc vào nó để bền vững.

Chính vì trí tuệ vừa là năng lực, vừa là đánh giá, nên bất kỳ định nghĩa "một câu" nào cũng dễ trở thành hẹp. Định nghĩa càng gọn, nó càng phải hy sinh nhiều thứ. Nếu ta định nghĩa trí tuệ là "khả năng giải quyết vấn đề", ta bỏ quên khả năng đặt vấn đề - một năng lực còn quan trọng hơn, vì câu hỏi sai thì lời giải đúng cũng vô nghĩa. Nếu ta định nghĩa trí tuệ là "khả năng học", ta lại phải hỏi: học cái gì, học đến mức nào, và học để làm gì - bởi một hệ thống có thể học rất nhanh những thói quen tệ hại. Nếu ta định nghĩa trí tuệ là "khả năng suy luận logic", ta sẽ đánh giá thấp những vùng không thể suy luận theo kiểu hình thức: thẩm mỹ, trực giác, bối cảnh xã hội, hay những quyết định phải đưa ra khi không đủ dữ liệu. Nếu ta định nghĩa trí tuệ là "khả năng tối ưu", ta sẽ vướng vào câu hỏi về mục tiêu: tối ưu cái gì, và cái gì bị bỏ qua vì khó đo. Những bẫy này khiến một cách tiếp cận mềm dẻo hơn trở nên hữu ích: thay vì cố đóng khung trí tuệ thành

một khối, hãy xem nó như một chùm năng lực đan xen, có thể mạnh yếu khác nhau ở từng người và từng loài. Một cá nhân có thể rất mạnh ở suy luận phân tích nhưng yếu ở giao tiếp xã hội; một cá nhân khác có thể không xuất sắc trong bài kiểm tra học thuật nhưng lại cực kỳ giỏi thích nghi, đọc tín hiệu môi trường, và đưa ra lựa chọn đúng thời điểm. Trí tuệ, theo nghĩa đó, không phải một chiếc cúp; nó là một bộ công cụ để tồn tại và định hướng trong thế giới thật.

Nếu ta chấp nhận trí tuệ như một tập hợp năng lực, ta sẽ bắt đầu thấy những lớp cấu thành của nó rõ hơn. Có lớp liên quan đến học từ kinh nghiệm: khả năng rút ra quy luật từ những lần thử - sai, rồi cập nhật hành vi để không lặp lại sai lầm cũ. Có lớp liên quan đến dự đoán và lập kế hoạch: không chỉ phản ứng với hiện tại, mà còn tưởng tượng được tương lai gần, cân nhắc hệ quả, và chọn đường đi phù hợp. Có lớp liên quan đến quản trị tài nguyên: thời gian, năng lượng, chú ý, tiền bạc - trí tuệ đôi khi là khả năng không tự đốt mình trong những cuộc đua vô nghĩa. Có lớp liên quan đến hiểu người khác: đọc ý định, đoán phản ứng, phối hợp trong nhóm - một năng lực sống còn của xã hội loài người. Có lớp liên quan đến ngôn ngữ: khả năng dùng biểu tượng để truyền ý, thuyết phục, và xây dựng tri thức chung. Và có một lớp sâu hơn mà ta ít nói nhưng luôn dùng: năng lực chọn mục tiêu. Bởi có những người suy luận tốt, học nhanh, làm việc hiệu quả - nhưng dùng toàn bộ năng lực ấy để theo đuổi một mục tiêu sai, hoặc mục tiêu do người khác "cài đặt" vào họ. Trong trường hợp đó, trí tuệ dường như không cứu họ; nó chỉ khiến họ chạy nhanh hơn về phía sai lầm.

Nhưng đến đây, một bước ngoặt quan trọng xuất hiện - và nó quyết định cách ta nhìn AI sau này. Nếu cứ nhìn trí tuệ bằng lăng kính con người, ta sẽ mắc kẹt trong thiên kiến của chính mình. Con người có một bản năng nhận dạng: ta công nhận trí tuệ khi thấy nó giống ta. Ta đánh giá cao lập luận bằng ngôn ngữ vì ngôn ngữ là vũ khí mạnh nhất của loài người. Ta dễ coi thường những dạng thông minh không thể diễn đạt thành câu chữ. Ta gán cái tên "bản năng" cho những hành vi cực kỳ tinh vi mà động vật thực hiện - chỉ vì chúng không viết luận văn về hành vi ấy. Ta gọi một loài là "kém thông minh" chỉ vì nó không làm tốt những bài kiểm tra do con người thiết kế, dù nó lại xuất

sắc trong môi trường sống của nó theo những cách chúng ta không thể bắt chước. Đây là cái bẫy "chuẩn vàng": lấy khả năng của mình làm chuẩn để đo mọi thứ khác, rồi kết luận rằng thứ khác kém vì không giống mình.

Một cách "giải độc" là thử đứng ra ngoài loài người trong tưởng tượng - như một kẻ quan sát ngoài hành tinh, không có lịch sử, không có tự tôn, không có lý do gì để tôn trí tuệ loài người làm trung tâm của vũ trụ. Bạn sẽ đánh giá trí tuệ bằng tiêu chuẩn nào? Bạn sẽ không hỏi "nó có nói được như con người không", bởi ngôn ngữ loài người khi ấy chỉ là một hiện tượng địa phương. Bạn cũng sẽ không hỏi "nó có giải được bài toán của con người không", bởi bài toán của con người phản ánh môi trường và văn hóa con người. Thay vào đó, bạn sẽ nhìn trí tuệ như dấu hiệu của một hệ thống biết tổ chức hành vi để đạt mục tiêu trong điều kiện hạn chế: biết học từ tín hiệu, biết thích nghi khi môi trường đổi, biết phối hợp khi lợi ích cần phối hợp, biết tối ưu hóa đường sống còn khi tài nguyên khan hiếm. Khi nhìn như vậy, bạn sẽ thấy trí tuệ có thể hiện diện trong nhiều hình dạng: một đàn sinh vật phối hợp như một cơ thể chung; một cá thể nhỏ bé nhưng biết dùng môi trường để tạo lợi thế; một hệ thống không có ngôn ngữ nhưng biết đo, biết dự đoán, biết lựa chọn hành động phù hợp.

Góc nhìn "ngoài hành tinh" ấy không chỉ là trò tưởng tượng; nó có tác dụng thực tế: nó buộc ta tách trí tuệ khỏi vẻ bề ngoài quen thuộc. Khi trí tuệ được hiểu như khả năng thích nghi và đạt mục tiêu, ta sẽ dễ chấp nhận rằng trí tuệ có thể tồn tại dưới những hình thức không giống con người - và đây chính là cánh cửa để hiểu AI một cách chín chắn. Bởi AI, nếu thành công, sẽ không nhất thiết là "một con người bằng silicon". Nó có thể là một dạng trí tuệ khác: mạnh ở tốc độ, mạnh ở quy mô, mạnh ở khả năng nhân bản; yếu ở thường thức, yếu ở bối cảnh, yếu ở hiểu ngầm xã hội; và đôi khi nguy hiểm không phải vì nó "xấu", mà vì nó tối ưu hóa mục tiêu theo cách lạnh lùng, nhất quán, và có thể được triển khai hàng loạt. Nếu ta cứ đòi AI phải "giống con người" mới gọi là trí tuệ, ta sẽ mù trước những năng lực thật của nó. Ngược lại, nếu ta hiểu trí tuệ rộng hơn, ta sẽ nhìn AI rõ hơn: ta sẽ hỏi nó đang học gì, thích nghi ra sao, tối ưu mục tiêu nào, và cái gì bị bỏ qua trong quá trình tối ưu ấy.

Và cuối cùng, có một lớp ý nghĩa sâu hơn: khi ta tranh luận trí tuệ là gì, ta đang tranh luận về chính con người. Ta đang nói điều gì đáng giá trong đời sống: tốc độ hay chiều sâu, hiệu quả hay ý nghĩa, tối ưu hay nhân văn, đúng theo luật hay đúng theo bối cảnh. Trí tuệ, vì thế, không chỉ là khái niệm khoa học; nó là chiếc gương phản chiếu giá trị. Chính vì nó vừa là năng lực vừa là giá trị, nên định nghĩa trí tuệ mới khó hơn ta tưởng và chính vì khó, nên nó đáng được đặt lại, trước khi ta giao ngày càng nhiều quyết định cho những hệ thống ngày càng "thông minh" theo cách riêng của chúng.

2. Trí tuệ không chỉ ở con người: bài học từ ong và nhện

Khi ta nói về trí tuệ, con người thường có thói quen nhìn vào chính mình rồi vô thức biến hình ảnh ấy thành chuẩn mực: trí tuệ được tưởng tượng như một thứ diễn ra trong đầu, được vận hành bằng ngôn ngữ, được chứng minh bằng lập luận, và được đo bằng những bài kiểm tra do con người đặt ra cho con người. Nhưng chỉ cần bước ra khỏi chiếc khung ấy và nhìn vào thế giới sống quanh ta, ta sẽ thấy một thực tế vừa giản dị vừa làm lung lay sự tự tin của chúng ta: trí tuệ không nhất thiết phải mang hình dạng "tư duy bằng lời". Có những dạng trí tuệ không cần câu chữ, không cần khái niệm trừu tượng, không cần triết lý tự sự; chúng không tồn tại như một bài diễn thuyết nội tâm, mà như một năng lực được đúc trực tiếp vào hành vi, được mài sắc bởi chọn lọc tự nhiên, và được tối ưu hóa theo môi trường sống đến mức tinh vi. Khi trí tuệ được nhìn theo cách này, nó không còn là một chiếc huy chương của loài người, mà là một phổ rộng các cơ chế "biết làm đúng" trong điều kiện thiếu chắc chắn: biết thu nhận tín hiệu từ môi trường, biết biến tín hiệu thành quyết định, biết phối hợp khi cần phối hợp, biết hành động có mục tiêu, biết điều chỉnh khi hoàn cảnh đổi thay. Thứ "biết" ấy có thể không được phát biểu, nhưng nó vẫn là tri thức theo nghĩa chức năng - tri thức giúp tồn tại và phát triển - và đôi khi, chính vì không cần lời nói, nó lại vận hành nhanh, gọn, và chính xác theo cách khiến ta phải tự hỏi: liệu có bao nhiêu trong cái ta gọi là "tư duy cao cấp" thực ra cũng là những cơ chế lựa chọn hành động, chỉ khác ở lớp vỏ ngôn ngữ và ý thức mà ta quen thuộc.

Ong là một ví dụ kinh điển khiến ta khó giữ thái độ kiêu ngạo. Một con ong mật sống trong một xã hội chặt chẽ, nơi hiệu quả của cả đàn phụ thuộc vào việc cá thể có thể biến trải nghiệm riêng thành thông tin chung. Khi một con ong bay đi tìm phấn hoa, nó không chỉ "thử vận may"; nó tương tác với ánh sáng, gió, mùi hương, địa hình, khoảng cách, và những ràng buộc về năng lượng. Nó phải cân bằng giữa bay xa để tìm nguồn tốt hơn và quay về kịp thời để không hao tổn quá mức. Rồi khi trở về, điều đáng kinh ngạc không nằm ở chỗ nó mang phấn về, mà nằm ở cách nó "kể lại" chuyến đi bằng hành vi: điệu nhảy của ong giống như một bản đồ được mã hóa. Hướng của

nguồn thức ăn được biểu đạt qua góc lệch so với hướng mặt trời, còn khoảng cách được biểu đạt qua độ dài và nhịp điệu của chuyển động; nghĩa là trong một chuỗi động tác tưởng như bản năng, có cả một hệ thống quy chiếu và một cơ chế truyền đạt. Nhưng điều sâu hơn là: đàn ong "đọc" được điệu nhảy ấy và biến nó thành hành động phối hợp. Thông tin được chuyển từ cá thể sang tập thể, từ tập thể thành luồng di chuyển, từ luồng di chuyển thành cơ hội sống còn. Nói cách khác, trí tuệ ở đây không nằm trong một cái đầu đơn lẻ mà nằm trong một mạng tương tác: một cá thể thu thập tín hiệu, mã hóa tín hiệu, cộng đồng giải mã tín hiệu, rồi cả hệ thống cập nhật hành vi. Đây là một dạng trí tuệ xã hội cực kỳ thực dụng: không cần ý niệm triết học, không cần "tự nhận thức" theo kiểu con người, nhưng vẫn có đầy đủ những thành phần mà ta thường gắn cho trí tuệ - học từ kinh nghiệm, truyền thông tin, tối ưu hóa hành vi theo mục tiêu, và tăng xác suất thành công trong môi trường bất định.

Nhện lại mở ra một cực khác của phổ trí tuệ: ít xã hội, nhiều chiến lược; ít truyền tin, nhiều kiến tạo. Ở nhện, trí tuệ không biểu lộ bằng giao tiếp tập thể, mà bằng khả năng biến môi trường thành lợi thế - bằng thiết kế cấu trúc, cảm nhận tín hiệu, và tối ưu chiến thuật săn mồi. Mạng nhện không phải là một tấm lưới ngẫu nhiên; nó là một công trình vừa là bẫy vừa là cảm biến: tơ nhện được sắp xếp theo cách phù hợp với loại con mồi, vị trí giăng, hướng gió, độ rung, độ dính, và cả "chi phí" mà nhện phải bỏ ra để tạo lại mạng khi bị phá. Khi một con mồi chạm vào, mạng truyền rung động như một kênh thông tin; nhện "đọc" rung động để ước lượng vị trí và mức độ mắc kẹt, rồi chọn thời điểm lao tới. Đặc biệt, nhện nước - loài sống trong ao hồ - cho thấy một kiểu trí tuệ kiến tạo gần như thơ mộng: nó dựng "chuông lặn" bằng tơ để tạo túi khí dưới nước, biến một môi trường vốn không dành cho hô hấp của nó thành nơi cư trú và phục kích. Đó là một chuỗi năng lực kết hợp: mang không khí xuống, duy trì cấu trúc, lựa chọn vị trí, điều chỉnh khi điều kiện thay đổi, rồi săn mồi khi cơ hội xuất hiện. Không cần câu chữ, vẫn có một năng lực tổ chức hành động để đạt mục tiêu; không cần "lý thuyết" theo nghĩa con người, vẫn có chiến thuật theo nghĩa sinh tồn. Và chính những ví dụ như ong và nhện nhắc ta rằng trí tuệ có nhiều hình thức, nhiều "kiến

trúc” và nhiều cách biểu lộ; từ đó, khi nói đến AI, câu hỏi đúng không còn là “máy có giống người không” - bởi giống người đôi khi chỉ là một lớp trình diễn thuyết phục - mà là máy có thể đạt được các năng lực cốt lõi của trí tuệ theo cách riêng của nó hay không, nó học từ tín hiệu nào, nó tối ưu mục tiêu ra sao, nó thất bại ở điểm nào, và khi năng lực ấy được nhân bản và triển khai ở quy mô lớn, xã hội sẽ làm gì để sức mạnh khuếch đại ấy phục vụ con người thay vì làm con người trở thành phân phụ của một cô máy tối ưu hóa.

3. Khi máy trở thành "thực thể trí tuệ": khác con người, nhưng không vì thế mà kém

Khi ta nói "máy trở thành một thực thể trí tuệ", điều đó không nhất thiết có nghĩa là máy đang tiến dần đến việc trở thành "một con người phiên bản kim loại", biết suy tư, biết tự sự, biết buồn vui theo cách ta quen thuộc; nó có thể đơn giản hơn mà cũng sâu sắc hơn: máy bắt đầu sở hữu những năng lực mà trước đây ta chỉ gán cho trí tuệ - năng lực nhận biết, dự đoán, lựa chọn, tối ưu, và đôi khi là hành động thay cho con người trong những điểm nút quan trọng của đời sống. Từ lâu, ta đã quen coi máy tính như một công cụ: nhập lệnh, xuất kết quả; dùng nó như một chiếc búa gõ nhanh hơn, một chiếc máy tính bỏ túi phóng đại khả năng tính toán của ta. Nhưng khi máy không chỉ thực hiện mệnh lệnh mà còn đề xuất mệnh lệnh; khi nó không chỉ trả lời câu hỏi mà còn gợi ra câu hỏi; khi nó không chỉ làm phép tính mà còn "đưa ra quyết định" thay ta trong việc sắp xếp tin tức, xếp hạng hồ sơ, phân loại rủi ro, gợi ý điều ta nên mua, nên xem, nên tin - thì vai trò của nó thay đổi từ công cụ thụ động sang một dạng tác nhân có ảnh hưởng. Và chính ở khoảnh khắc ấy, ngôn ngữ của chúng ta cũng bắt đầu thay đổi: ta nói máy "thông minh", máy "hiểu", máy "biết", máy "nhận ra", như thể nó đang bước vào vùng lãnh địa vốn thuộc về con người. Nhưng đây cũng là lúc cần tỉnh táo: một thực thể trí tuệ không nhất thiết phải giống ta mới có sức mạnh; thậm chí, chính sự khác biệt của nó mới làm nên dạng quyền lực mới - quyền lực của tốc độ, của quy mô, của sự bền bỉ, của khả năng lặp lại và nhân bản.

Máy tính có những lợi thế khiến con người khó cạnh tranh nếu đặt lên bàn cân thẳng thừng: tốc độ xử lý có thể tăng đến mức gần như "tức thì" ở tầm trải nghiệm con người; độ chính xác tính toán ở những miền xác định gần như tuyệt đối; dung lượng ghi nhớ không chỉ lớn mà còn có thể truy hồi theo cấu trúc; khả năng vận hành liên tục không cần ngủ, không cần nghỉ, không có cảm xúc làm chệch nhịp; và quan trọng hơn, nó có thể được sao chép - một năng lực không chỉ mạnh mà còn có thể được triển khai đồng loạt. Một bác sĩ giỏi có thể cứu được nhiều người, nhưng vẫn bị giới hạn bởi thời gian, sức khỏe và đời sống cá nhân; một hệ thống dự đoán giỏi có thể được nhân lên thành hàng nghìn bản, chạy song song trên hạ tầng, phục vụ hàng

triệu quyết định trong cùng một khoảng thời gian. Chính vì thế, khi máy đạt tới mức đủ hữu ích, nó không còn là "một công cụ tốt", mà trở thành một "hạ tầng" vô hình: nền cho việc ra quyết định, nền cho việc phân phối sự chú ý, nền cho việc định hướng lựa chọn. Và khi trí tuệ trở thành hạ tầng, sự khác biệt giữa con người và máy không còn nằm ở chỗ ai "thông minh" hơn theo nghĩa tuyệt đối, mà nằm ở chỗ ai có lợi thế về tốc độ, quy mô và khả năng khuếch đại - những thứ có thể thay đổi kết cấu xã hội nhanh hơn ta kịp thích nghi.

Nhưng chính cái "khác" của máy lại là nguồn gây hiểu nhầm lớn nhất: ta dễ tưởng máy "hiểu" vì máy "làm đúng", giống như ta dễ gán cho một người nói trôi chảy là người hiểu sâu. Trong đời sống thường ngày, con người dùng rất nhiều tín hiệu bề mặt để đánh giá: giọng nói tự tin, câu chữ mạch lạc, phản hồi nhanh, dẫn chứng hợp lý, thái độ điềm tĩnh - những tín hiệu ấy thường tương quan với năng lực, nhưng không bảo đảm năng lực. Máy, đặc biệt là các hệ thống có khả năng tạo ngôn ngữ hoặc đưa ra dự đoán dựa trên mẫu hình, có thể khai thác chính thói quen đánh giá ấy của chúng ta. Nó có thể trả lời đúng theo những cách không giống người - đúng nhờ thông kê và tối ưu, đúng nhờ khớp mẫu hình, đúng nhờ học từ dữ liệu khổng lồ - và vì đúng, ta suy rộng ra rằng nó hiểu. Nhưng hiểu theo nghĩa con người thường gắn với bối cảnh, với mục đích, với khả năng giải thích vì sao, với sự tự kiểm tra khi không chắc chắn; còn máy có thể đưa ra kết quả đúng mà không có bất kỳ "trải nghiệm hiểu" nào như ta tưởng. Nguy hiểm hơn, máy cũng có thể sai theo những cách rất khó chịu: sai nhưng trôi chảy, sai mà tự tin, sai mà vẫn "có vẻ hợp lý", sai một cách nhất quán đến mức người dùng không nhận ra. Đó là kiểu sai khiến con người mất cảnh giác, bởi nó không mang dáng vẻ lúng túng; nó giống như một câu trả lời "chắc nịch" của một người nói giỏi và lịch sử con người cho thấy: những sai lầm nguy hiểm nhất thường không đến từ sự ngu dốt công khai, mà đến từ sự tự tin mượt mà.

Ở đây, ta chạm vào một điểm cốt lõi: trí tuệ máy không nhất thiết là phiên bản nhân bản của trí tuệ người. Nhiều hướng phát triển AI cổ điển từng cố mô phỏng "cách con người làm": viết luật, mã hóa tri thức, dựng mô hình suy luận, mô phỏng chuyên gia trong một miền hẹp; tham vọng của nó là làm cho máy suy nghĩ như một bộ óc lý

tưởng hóa - có cấu trúc, có suy diễn, có giải thích. Nhưng càng đi sâu, người ta càng gặp một sự thật: con người không luôn suy nghĩ theo cách mà ta nghĩ con người suy nghĩ. Ta có trực giác, có thiên kiến, có lối tắt tâm lý; ta dựa vào kinh nghiệm, vào cảm giác bối cảnh; ta đôi khi quyết định trước rồi mới tìm lý do sau. Mô phỏng "con người lý trí" hóa ra không phải mô phỏng con người thật. Và khi dữ liệu lớn cùng sức mạnh tính toán bùng nổ, một con đường khác trở nên hiệu quả hơn: không cố bắt máy đi theo đường của con người, mà để máy đi đường của máy - học từ dữ liệu, tối ưu theo mục tiêu, tìm ra mẫu hình mà con người không dễ nhìn thấy. Từ đó, một sự chấp nhận dần hình thành: máy có thể không giống người, miễn là nó hữu ích; nó có thể không suy luận như ta, miễn là nó đạt kết quả tốt và ổn định; nó có thể không "hiểu" theo nghĩa triết học, miễn là nó đáng tin trong phạm vi ứng dụng. Nhưng chữ "miễn là" ở đây cực kỳ quan trọng, bởi nó chứa toàn bộ trách nhiệm của thời đại: hữu ích phải đi kèm giới hạn rõ; đáng tin phải đi kèm kiểm chứng; ổn định phải đi kèm cơ chế phát hiện khi hệ thống gặp tình huống ngoài dự đoán; và nhất là, phải biết phân biệt giữa năng lực và quyền lực - vì một hệ thông có năng lực càng lớn, quyền lực quyết định càng lớn, thì cái giá của một sai lệch nhỏ càng có thể biến thành một sai lệch mang tính hệ thống.

Vì vậy, câu hỏi đúng khi máy trở thành "thực thể trí tuệ" không phải là hỏi nó có giống con người không, cũng không phải là vội vàng kết luận nó kém hay hơn; câu hỏi đúng là: máy đang mạnh ở đâu, yếu ở đâu, và sự mạnh - yếu ấy sẽ tái cấu trúc đời sống ra sao khi được nhân bản và cắm vào các quy trình xã hội. Con người mạnh ở hiểu ngầm bối cảnh, ở đạo đức tình huống, ở khả năng cảm nhận hậu quả xã hội, ở sự linh hoạt khi mục tiêu thay đổi; máy mạnh ở tốc độ, ở quy mô, ở khả năng lặp lại, ở tối ưu hóa nhất quán theo mục tiêu được đặt ra. Nếu mục tiêu đặt đúng, máy là đòn bẩy phi thường; nếu mục tiêu đặt sai, máy là máy khuếch đại sai lầm. Nếu dữ liệu đại diện, máy là công cụ mở rộng năng lực; nếu dữ liệu lệch, máy là công cụ hợp thức hóa thiên kiến bằng vẻ ngoài "khách quan". Và trong mọi trường hợp, điều khiến máy trở nên khác biệt không phải là nó có "linh hồn" hay không, mà là nó có khả năng biến một lựa chọn thành một chuẩn mực với tốc độ vượt xa tốc độ tự sửa sai của xã hội. Chỉ khi hiểu như

vậy, ta mới có thể bước vào kỷ nguyên trí tuệ nhân tạo với một thái độ trưởng thành: không phủ nhận, không sùng bái, mà biết đặt công nghệ vào đúng chỗ - một thực thể có năng lực đặc thù, cần được kiểm soát bằng thiết kế mục tiêu, cơ chế kiểm chứng và trách nhiệm xã hội, để sự khác biệt của nó trở thành lợi ích, thay vì trở thành khoảng trống nơi sai lầm được nhân bản vô hạn.

4. AI cổ điển: trí tuệ như biểu tượng, luật lệ và cấu trúc tri thức

AI cổ điển là một trong những nỗ lực nghiêm túc đầu tiên của con người nhằm biến "trí tuệ" thành một thứ có thể thiết kế được, lắp ráp được và kiểm soát được. Ở thời điểm mà dữ liệu chưa bùng nổ, sức mạnh tính toán còn hạn chế, và khái niệm "học từ dữ liệu khổng lồ" chưa trở thành dòng chảy chính, các nhà nghiên cứu đã đi theo một trực giác rất con người: nếu trí tuệ của chúng ta đến từ việc hiểu thế giới bằng khái niệm, lưu trữ tri thức thành những mảnh có cấu trúc, rồi suy luận theo các bước logic để đi từ cái biết tới cái chưa biết, thì máy cũng có thể thông minh nếu ta cho nó những thứ tương tự. Vì thế, AI cổ điển thường được hình dung như một công trình kiến trúc của biểu tượng và luật lệ: thế giới được "dịch" thành các ký hiệu, các mối quan hệ được viết thành cấu trúc, còn suy luận là một chuỗi thao tác có thể theo dõi. Từ góc nhìn này, trí tuệ không phải là thứ mơ hồ như cảm hứng; nó giống một hệ thống có thể mô tả, có thể phân rã, có thể định nghĩa. Khi ta nhìn vào một hành vi "thông minh" của con người - chẩn đoán bệnh, lập kế hoạch, tư vấn kỹ thuật, giải câu đố logic - ta thường thấy ở đó một vệt của kiến thức có cấu trúc và một chuỗi suy luận hợp lý, và AI cổ điển đặt cược rằng nếu ta mô hình hóa được vệt ấy, ta có thể tạo ra trí tuệ nhân tạo theo nghĩa máy có thể "lập luận" thay ta trong những miền có quy tắc.

Nhưng để làm được điều đó, AI cổ điển buộc phải trả lời hai câu hỏi nền tảng - hai câu hỏi nghe như triết học, nhưng lại đẻ ra cả một ngành kỹ thuật. Câu hỏi thứ nhất nằm ở phía tri thức: con người, đặc biệt là chuyên gia, đang "giữ" tri thức trong đầu theo kiểu gì mà họ có thể ra quyết định nhanh và chắc, đôi khi còn giải thích được vì sao; tri thức ấy có được tổ chức theo dạng danh sách, theo dạng phân loại, theo dạng nguyên tắc, hay theo dạng những "kịch bản" quen thuộc; và quan trọng hơn, làm sao biến một thứ vốn rất giàu bối cảnh và kinh nghiệm thành một mô hình đủ rõ để máy có thể lưu trữ và thao tác. Câu hỏi thứ hai nằm ở phía suy luận: một khi tri thức đã được biểu diễn, máy nên suy luận theo quy tắc nào để hành vi của nó giống "hành vi có trí tuệ" - tức không chỉ đúng trong một vài trường hợp mẫu, mà còn nhất quán, có thể dự đoán, và có thể giải thích. Hai câu

hỏi này thực ra kéo theo một nhiệm vụ khó: phải "đóng gói" thế giới thành những thứ có thể viết được, và phải "đóng gói" suy nghĩ thành những bước có thể chạy được. Đó là lý do AI cổ điển giống như nghệ thuật xây dựng một thành phố bằng bản vẽ: muốn có một con đường đi từ A đến B, bạn phải có bản đồ, phải có quy hoạch, phải có luật giao thông và tất cả đều phải được viết ra rõ ràng trước khi chiếc xe đầu tiên lăn bánh.

Từ chương trình ây, những công cụ tô chức tri thức ra đời như các "khung" (frames), nơi mỗi sự vật hay khái niệm được mô tả bằng một tập thuộc tính, quan hệ và các giá trị mặc định, có thể kế thừa theo tầng lớp giống như cách con người nghĩ về thế giới theo nhóm và loại. Một "chiếc xe" có những thuộc tính cơ bản; "xe cứu thương" kế thừa thuộc tính của "xe" nhưng có thêm thuộc tính riêng; "xe cứu thương trong bệnh viện X" lại có ngữ cảnh khác. Chính kiểu biểu diễn này cố gắng bắt chước một điều rất đời thường: chúng ta không học từng vật từ đầu; chúng ta học theo mẫu, theo lớp, theo khung tham chiếu. Cùng với đó, các hệ luật–quy tắc kiểu "nếu… thì…" trở thành xương sống của nhiêu hệ thông chuyên gia: nêu có triệu chứng A và B thì nghi ngờ C; nếu kiểm tra D ra E thì chuyển sang phác đồ F; nếu điều kiện G thỏa mãn thì đề xuất phương án H. Những luật này, khi được gom lại, trông giống như "trí tuệ đóng chai": một phần kinh nghiệm chuyên gia được trích xuất, chuẩn hóa và biến thành hệ thống. Nhưng đời sống không chỉ là luật; đời sống còn là lựa chọn giữa vô số khả năng, nên AI cổ điển cũng phát triển các kỹ thuật tìm kiếm lời giải trong không gian trạng thái: máy có thể phải thử nhiều đường đi để tìm ra đường tốt, giống như giải một mê cung. Vì thế xuất hiện những chiến lược "tìm tốt trước" (ưu tiên những bước hứa hẹn), các phương pháp leo đôi (cải thiện dần theo tiêu chí), tôi ưu cục bộ và các biên thê - tất cả nhằm mục tiêu giúp máy không bị chìm trong vô hạn lựa chọn, mà biết đi theo các dấu hiệu hướng tới mục tiêu nhanh hơn. Những cơ chế này khiến AI cổ điển không chỉ là "một bộ luật", mà là "một bộ luật biết tìm đường" trong một thế giới được mô hình hóa.

Điểm quyến rũ nhất của AI cổ điển là cảm giác an tâm mà nó mang lại cho con người: tính "giải thích được" tương đối. Khi hệ thống đưa ra quyết định, ta có thể lần theo dấu vết: luật nào đã kịch

hoạt, điều kiện nào đã thỏa, thuộc tính nào trong khung tri thức đã được dùng, và chuỗi suy luận nào dẫn đến kết luận. Trong những miền mà sự giải trình quan trọng - y tế, pháp lý, an toàn, tài chính - khả năng "nói ra vì sao" không phải phụ kiện; nó là điều kiện để xã hội chấp nhận. Nhưng chính sự rõ ràng ấy cũng làm lộ giới hạn của AI cổ điển: thế giới thật không vận hành gọn gàng như một cuốn sổ tay quy tắc. Tri thức ngoài đời thường thiếu, không đầy đủ; nhiều thông tin mơ hồ, có ngoại lệ; các quy tắc có thể mâu thuẫn tùy bối cảnh; và những thứ quan trọng nhất lại nằm trong "thường thức" mà con người hiếm khi nói ra vì tưởng ai cũng biết. Thường thức không giống một định luật vật lý có thể viết thành câu "nếu… thì…"; nó giống một mạng liên tưởng mềm dẻo, phụ thuộc vào ngữ cảnh, và thay đổi theo văn hóa. Khi cố viết thường thức thành quy tắc, ta sẽ gặp "bệnh ngoại lệ": quy tắc A đúng, nhưng có ngoại lệ B; thêm B thì lại phát sinh ngoại lệ C; và cứ thế hệ thống phình to, khó bảo trì, khó kiểm soát. Một bộ luật càng lớn càng dễ chứa xung đột, và càng khó chắc rằng nó bao phủ được những tình huống lạ của đời sống. AI cổ điển vì vậy mang một nghịch lý: muốn giải thích được thì phải mô tả thế giới rõ, nhưng mô tả càng rõ lại càng thấy thế giới không rõ như ta tưởng; muốn luật hóa thế giới để máy suy luận, lại càng thấy đời sống có quá nhiều điều "không luật nào ôm hết".

Tuy nhiên, sẽ là thiếu công bằng nếu coi AI cổ điển chỉ như một giai đoạn đã qua. Di sản quan trọng của nó nằm ở chỗ nó buộc chúng ta đối diện với hai câu hỏi mà AI hiện đại đôi khi lảng tránh: tri thức là gì và suy luận là gì. Dù hôm nay các hệ thống học sâu có thể đạt kết quả ấn tượng nhờ dữ liệu và quy mô, nhu cầu về cấu trúc, về giải trình, về suy luận có kiểm soát vẫn còn nguyên - đặc biệt khi AI được đưa vào những quyết định ảnh hưởng đời sống con người. AI cổ điển nhắc ta rằng trí tuệ không chỉ là dự đoán đúng; trí tuệ còn là biết vì sao mình đúng, biết khi nào mình không chắc, biết giới hạn của mình, và có thể trình bày đường đi của quyết định để người khác kiểm chứng. Và chính từ điểm yếu của AI cổ điển - khó nắm bắt thường thức, khó bao phủ ngoại lệ - ta cũng hiểu vì sao AI hiền đại trỗi dậy: nó chọn cách học từ thế giới thay vì cố viết lại thế giới. Nhưng để bước vào thời đại AI một cách trưởng thành, ta cần giữ lại tinh thần

của AI cổ điển như một phần kỷ luật trí tuệ: không bị mê hoặc bởi kết quả, luôn hỏi về cấu trúc tri thức, về quy tắc ra quyết định, và về khả năng giải thích - bởi trong xã hội, một hệ thống "đúng" mà không thể giải trình vẫn có thể là một hệ thống nguy hiểm.

5. Khi AI bước vào dữ liệu: khai phá mẫu hình và dự đoánh hành vi

Khi AI bước vào kỷ nguyên dữ liệu, nó bước vào bằng một cánh cửa khác hẳn so với AI cổ điển: không còn là tham vọng viết ra một bộ luật để máy suy luận giống con người, mà là tham vọng để máy "nhìn" thế giới theo dấu vết mà con người để lại, rồi tự rút ra những quy luật ẩn bên dưới. Thế giới hiện đại tạo ra dữ liệu như hơi thở: mỗi cú nhấp chuột, mỗi lượt xem, mỗi lần mua hàng, mỗi lần di chuyển, mỗi cuộc gọi, mỗi giao dịch, mỗi biểu mẫu đăng ký, mỗi lần dừng lại lâu hơn trước một bức ảnh - tất cả là các tín hiệu nhỏ gom lại thành một bức tranh lớn về hành vi. Trong bức tranh ấy, nhiều điều không cần ai kể, không cần ai giải thích bằng lời; chúng tự bộc lộ qua tần suất, qua chuỗi thời gian, qua sự đồng xuất hiện, qua nhịp điệu lặp lại, qua những lệch chuẩn bất thường. Và ở đây, "AI theo nghĩa thực dụng" có một nhiệm vụ rất rõ: thay vì ép máy suy luận như một nhà logic học, ta để máy làm một việc mà nó đặc biệt giỏi - đếm, đo, so sánh, phát hiện mẫu hình - rồi dùng các mẫu hình ấy để dự đoán, gợi ý, tối ưu. Khai phá dữ liệu vì thế không chỉ là một kỹ thuật; nó là một cách nhìn đời sống như một hệ thống tín hiệu: ai thường mua gì vào thời điểm nào, họ mua kèm thứ gì, họ quay lại sau bao lâu, họ rời bỏ vì lý do gì; thị trường biến động ra sao, nhịp cầu cung - cầu đổi thế nào; hành vi nào là bình thường của một nhóm, hành vi nào là lệch chuẩn báo hiệu rủi ro. Khi dữ liệu đủ nhiều, một điều kỳ lạ xảy ra: những điều ta tưởng là ngẫu nhiên bắt đầu có cấu trúc; và những điều ta tưởng là "tự nhiên" của cá nhân bắt đầu hiện ra như các mẫu hành vi của tập thể.

Điểm hấp dẫn của khai phá dữ liệu nằm ở chỗ nó có thể biến "đám đông mơ hồ" thành những nhóm có chân dung tương đối rõ, và biến "cảm giác" của nhà quản lý thành những dự đoán có xác suất. Trong tiếp thị và bán lẻ, dữ liệu trở thành chiếc kính phóng đại hành vi: thay vì nói chung chung rằng "khách hàng thích giảm giá", doanh nghiệp có thể phân nhóm theo thói quen tiêu dùng, theo mức độ trung thành, theo độ nhạy cảm với giá, theo sở thích theo mùa, theo kênh tương tác, theo bối cảnh địa lý. Một khách hàng mua đều đặn mỗi tháng là một câu chuyện khác với khách hàng chỉ xuất hiện khi có khuyến mãi

người mua quà tặng theo dịp là một câu chuyện khác với người mua cho nhu cầu sử dụng hàng ngày; thường mua kèm hai món là một câu chuyện khác với người chỉ mua một món rồi rời đi. Từ đó, hệ thống có thể đưa ưu đãi "đúng người - đúng lúc - đúng kênh" thay vì bắn đại trà. Nhưng điều thú vị không chỉ là nhắm mục tiêu; điều sâu hơn là tối ưu trải nghiệm: gợi ý sản phẩm tiếp theo, dự đoán khả năng rời bỏ để kịp thời giữ chân, thiết kế thông điệp phù hợp với tâm thế của nhóm. Và khi hệ thống vận hành đủ lâu, nó trở thành một vòng lặp: dữ liệu tạo dự đoán, dự đoán tạo hành động (gợi ý, ưu đãi, hiển thị), hành động tạo hành vi mới, hành vi mới lại xem như bằng chứng để hệ thống học tiếp. Vòng lặp này là nguồn sức manh, nhưng cũng là nơi dễ sinh ra những hiệu ứng phụ mà nếu không nhìn kỹ, ta sẽ tưởng rằng "thị trường tự nhiên vốn vậy", trong khi thực ra hệ thống đã góp phần tạo ra thị trường theo cách nó tối ưu.

Trong kinh doanh và tài chính, khai phá dữ liệu thường mang dáng vẻ nghiêm túc và lạnh lùng hơn, bởi nó động đến dự báo, rủi ro và ra quyết định. Dữ liệu ở đây không chỉ kể chuyện quá khứ; nó cố gắng nói về tương lai: xu hướng tiêu dùng sẽ đi về đâu, nhu câu sẽ tăng hay giảm, nguồn cung có bị đứt gãy không, giá sẽ biến động theo kịch bản nào, khả năng trả nợ của một nhóm khách hàng ra sao, mức độ rủi ro của một khoản vay là bao nhiêu. Những hệ thống dự báo không thay thế hoàn toàn con người, nhưng chúng thay đổi nhịp độ ra quyết định: thay vì họp bàn dài ngày để tranh luận bằng trực giác, người ta có thể nhìn vào các mô hình xác suất, các chỉ số rủi ro, các mô phỏng kịch bản. Lợi ích rõ ràng: ra quyết định nhanh hơn, nhất quán hơn, có thể đo được hiệu quả. Nhưng cùng lúc, một thứ khác cũng tăng lên: quyền lực của mô hình. Khi mô hình được tích hợp vào quy trình, nó không chỉ là "tham khảo"; nó dần trở thành tiêu chuẩn vận hành - ai được vay, ai bị từ chối, ai bị đánh giá rủi ro cao, ai phải chịu lãi suất khác. Và khi tiêu chuẩn ấy đã được tự động hóa, sai lệch không còn là sai lệch của một cá nhân; nó có thể trở thành sai lệch của cả hệ thống. Chính vì vậy, khai phá dữ liệu trong tài chính không chỉ là bài toán kỹ thuật, mà còn là bài toán công bằng, giải trình và trách nhiệm: mô hình dựa trên dữ liệu nào, dữ liệu có đại diện không, và ai chịu trách nhiệm khi dự đoán sai gây tổn hại thật cho đời sống con người.

Trong an ninh và chống gian lận, khai phá dữ liệu lại thể hiện một vai trò khác: phát hiện những điều "không giống bình thường". Ở đây, trọng tâm không phải là gợi ý mua thêm, mà là gắn cờ rủi ro: giao dịch này có khác hành vi trước đây không, thời điểm này có bất thường không, địa điểm này có hợp lý không, tốc độ và chuỗi hành động có dấu hiệu của kịch bản lừa đảo không. Một người thường mua sắm trong một khu vực, bỗng xuất hiện giao dịch ở nơi xa; một thẻ thường chi tiêu theo nhịp nhỏ, bông có khoản lớn liên tiếp; một tài khoản thường hoạt động ban ngày, bỗng đăng nhập dồn dập lúc rạng sáng; một người dùng bình thường gõ và nhấp theo nhịp tự nhiên, bỗng hành vi giống bot - tất cả là những tín hiệu mà hệ thống có thể phát hiện nhanh hơn con người. Đây là nơi AI "thực dụng" tỏa sáng: nó không cần hiểu câu chuyện đạo đức, nó chỉ cần nhận ra sai lệch thống kê và báo động. Nhưng chính vì thế, nó cũng có nguy cơ gây hại theo cách âm thầm: báo động sai khiến người vô tội bị chặn giao dịch, bị nghi ngờ, bị làm khó; báo động quá nhiều dẫn tới "chai lì cảnh báo"; báo động quá ít khiến kẻ xấu lọt qua. Và như mọi hệ thống an ninh khác, đây là cuộc chơi hai phía: kẻ gian lận quan sát mô hình, tìm cách "trông giống bình thường", lách qua ngưỡng phát hiện; hệ thống lại phải cập nhật, siết, hoặc thay đổi tiêu chí. Khi dữ liệu lớn hơn, cuộc chơi này nhanh hơn, và cái giá của sai sót cũng lớn hơn, bởi mọi thứ xảy ra trong thời gian thực với quy mô lớn.

Điều cần giữ trong đầu, để không bị thôi miên bởi sức mạnh của dữ liệu, là: mô hình càng mạnh khi dữ liệu càng lớn, nhưng cũng càng nguy hiểm nếu ta nhầm lẫn điều gì đang thực sự được học và điều gì chỉ là ảo ảnh của tương quan. Dữ liệu có thể cho thấy hai sự kiện thường đi cùng nhau, nhưng không tự động trả lời câu hỏi vì sao; tương quan không phải nguyên nhân, và nếu ta dùng mô hình tương quan để hành xử như thể mình đã nắm được nguyên nhân, ta có thể tạo ra những quyết định sai mà vẫn "có vẻ khoa học". Nguy hiểm hơn nữa là khi dự đoán trở thành mệnh lệnh: khi một điểm số rủi ro không còn là gợi ý mà trở thành quyết định tự động; khi một dự đoán về "khả năng bỏ học" hay "khả năng phạm lỗi" không còn là cảnh báo để hỗ trợ mà trở thành cái nhãn dán định mệnh lên một con người. Ở đó, mô hình không chỉ phản ánh thế giới, nó góp phần tạo ra thế giới: bị

chấm điểm thấp thì cơ hội ít đi, cơ hội ít đi thì kết quả xấu đi, kết quả xấu đi lại "xác nhận" mô hình; một vòng lặp tự củng cố hình thành, biến dự đoán thành lời tiên tri tự hoàn thành. Vì vậy, khai phá dữ liệu và dự đoán hành vi phải đi kèm một kỷ luật trưởng thành: phải có cơ chế kiểm chứng độc lập, phải có khả năng giải trình ở mức phù hợp, phải có giám sát liên tục khi bối cảnh thay đổi, và phải có đường lui khi hệ thống sai - bởi trong đời sống con người, một dự đoán sai không chỉ là một con số sai; nó có thể là một cánh cửa bị đóng lại, một cơ hội bị tước đi, một niềm tin bị phá vỡ, và đôi khi là một vết hằn dài trong số phận của một người không hề biết mình đã bị một mô hình "đọc" như thế nào.

6. AI và triết học: máy có "hiểu" hay chỉ "đóng vai hiểu"?

Ngay khi máy bắt đầu làm được những việc từng được coi như "dấu hiệu nhận dạng" của trí tuệ - giải bài toán phức tạp, chơi cờ ở đẳng cấp cao, nhận diện hình ảnh, dịch ngôn ngữ, trò chuyện trôi chảy - một cánh cửa khác lập tức mở ra, không phải cánh cửa kỹ thuật mà là cánh cửa triết học: nếu máy có thể làm những điều mà ta từng nghĩ chỉ trí tuệ mới làm được, vậy máy có đang "nghĩ" hay không, hay nó chỉ đang thực hiện một chuỗi thao tác tinh vi đến mức tạo ra ảo giác về tư duy. Câu hỏi này khiến nhiều người bối rối vì nó chạm vào một vùng mờ giữa hành vi và bản chất: ta có thể quan sát hành vi của máy, có thể đo hiệu suất, có thể so sánh kết quả với con người, nhưng ta không có cách trực tiếp để nhìn thấy "bên trong" theo kiểu ta vẫn tưởng mình nhìn thấy bên trong chính mình. Và vì triết học vốn không chịu dừng lại ở câu trả lời tiện lợi, nó buộc ta phải đặt lại những điều tưởng như chắc chắn: "hiểu" nghĩa là gì, "nghĩ" nghĩa là gì, và ta đang dùng tiêu chuẩn nào để gắn nhãn trí tuệ cho một thực thể khác. Ở đây, điều khó nhất không phải là máy, mà là con người: chúng ta mang theo một trực giác sâu trong xương tủy rằng trí tuệ gắn với trải nghiệm chủ quan, với ý thức, với cảm nhận; và khi ta gặp một thứ "làm được" mà không chắc "có cảm nhận" hay không, ta rơi vào thế lưỡng nan - hoặc ta mở rộng định nghĩa trí tuệ, hoặc ta giữ định nghĩa cũ và nói rằng máy chỉ đang bắt chước. Cả hai lựa chọn đều có cái giá của chúng: mở rộng quá dễ thì ta có thể sùng bái công cụ; giữ chặt quá lâu thì ta có thể tự bịt mắt trước một dạng quyền lực mới.

Chính vì vậy tranh luận quanh trí tuệ máy thường trở nên gay gắt và đôi khi cảm tính, bởi nó không chỉ là tranh luận về công nghệ mà còn là tranh luận về vị trí của con người. Có một phản ứng rất quen thuộc lặp đi lặp lại qua lịch sử: mỗi khi máy làm được một việc mà trước đó ta tự hào là "đặc quyền" của con người, ta lập tức dời cột mốc và nói: "máy giỏi đấy, nhưng nó không thể…". Nó không thể sáng tạo thật; không thể hiểu thật; không thể có cảm xúc thật; không thể có đạo đức thật; không thể có ý thức thật. Chuỗi "không thể" ấy vừa là nỗ lực bảo vệ giá trị con người, vừa là dấu hiệu của một nỗi lo sâu hơn: nếu ta chấp nhận rằng một thực thể không giống ta vẫn có thể thực hiện những chức năng trí tuệ, ta sẽ buộc phải đối diện với câu

hỏi về bản chất con người, về điều gì thực sự làm nên "tính người" ngoài năng lực xử lý. Điều đáng lưu ý là phản ứng tự vệ này không hẳn sai; nó có thể là một lời nhắc rằng hành vi bề ngoài không đủ để kết luận về chiều sâu nội tại. Nhưng nó cũng có thể trở thành cái bẫy: ta cứ liên tục dời tiêu chuẩn, và cuối cùng biến trí tuệ thành một khái niệm "không bao giờ đạt tới" đối với bất cứ thứ gì không phải con người - một kiểu đặc quyền định nghĩa, nơi con người vừa là giám khảo vừa là thí sinh duy nhất được phép thắng. Khi đó, tranh luận không còn nhằm hiểu sự thật, mà nhằm bảo vệ một cảm giác ưu thế, và chính cảm giác ấy lại dễ khiến ta chủ quan trước những hệ thống đang ảnh hưởng thật đến đời sống xã hội.

Trong nỗ lực tìm một thước đo "khách quan" cho trí tuệ máy, người ta thường nhắc đến một phép thử nổi tiếng: trò "bắt chước" trong hội thoại, thứ mà công chúng hay gọi là bài kiểm tra Turing. Ý tưởng của nó có sức hấp dẫn giản dị: nếu bạn không thể phân biệt một cách đáng tin giữa câu trả lời của máy và câu trả lời của người trong một cuộc trò chuyện, thì ít nhất máy đã đạt được một dạng năng lực giao tiếp mà ta thường gắn với trí tuệ. Nhưng sức hấp dẫn ấy đồng thời che giấu một sự tinh tế: bài kiểm tra này không đo trực tiếp "bản chất của hiểu biết", mà đo khả năng tạo ra hành vi giao tiếp đủ thuyết phục trong một bối cảnh nhất định. Và khi máy tiến gần đến khả năng đối thoại tự nhiên, câu hỏi lớn không biến mất mà lại trở nên sắc hơn: vượt qua một bài kiểm tra xã hội có đồng nghĩa với việc sở hữu trí tuệ như con người không, hay nó chỉ chứng minh rằng máy đủ giỏi trong việc tạo ra ấn tượng về trí tuệ. Nếu máy "đánh lừa" được ta, điều đó nói gì về máy - rằng máy có năng lực ngôn ngữ mạnh, có khả năng mô phỏng phong cách, có thể điều chỉnh theo ngữ cảnh - nhưng đồng thời cũng nói gì về chính con người: rằng ta thường đánh giá trí tuệ dựa trên bề mặt của ngôn từ, dựa trên sự mạch lạc và tự tin, và rằng nhiều khi cái ta gọi là "hiểu" trong giao tiếp hàng ngày thực ra là một cảm giác được tạo bởi sự trôi chảy của câu chữ. Khi ấy, bài kiểm tra không chỉ là thử thách cho máy, mà còn là tấm gương soi thói quen nhận thức của con người.

Thực tế, đã có những tình huống cho thấy người đánh giá đôi khi nhầm lẫn đáng kể khi phân biệt máy với người trong các cuộc đối

thoại dạng kiểm tra, đặc biệt khi bối cảnh bị giới hạn và người tham gia có kỳ vọng nhất định. Nhưng ngay cả nếu chuyện nhầm lẫn xảy ra thường xuyên hơn, tranh cãi vẫn không chấm dứt, bởi vấn đề nằm ở chỗ ra đang đo "trình diễn giống người", thì một hệ thống có thể tối ưu hóa để diễn tốt mà không cần sở hữu bất kỳ chiều sâu nào gọi là hiểu; nó có thể học cách né câu hỏi khó, học cách nói chung chung, học cách phản chiếu cảm xúc của người đối thoại, học cách tạo ấn tượng về sự nhất quán. Ngược lại, nếu ta muốn đo "bản chất của hiểu biết", ta phải trả lời một câu hỏi còn khó hơn: hiểu biết là gì, và ta có thể kiểm chứng hiểu biết bằng phương pháp nào mà không dựa vào chính các biểu hiện bên ngoài. Bởi ngay cả giữa con người với nhau, ta cũng không chạm trực tiếp được vào ý thức của người khác; ta chỉ suy ra từ hành vi, lời nói, và sự nhất quán theo thời gian. Vì thế, tranh luận về trí tuệ máy thường bị mắc kẹt giữa hai cực: một cực nhấn mạnh hành vi và tính thực dụng - nếu nó làm được thì nó "có trí tuệ" theo nghĩa chức năng; cực kia nhấn mạnh trải nghiệm nội tại và tính bản thể - nếu không có ý thức hay hiểu biết theo nghĩa sâu thì nó chỉ là bắt chước. Sự thật có thể nằm ở việc ta cần nhiều lớp tiêu chuẩn hơn: phân biệt trí tuệ như năng lực hành động trong môi trường, trí tuệ như năng lực ngôn ngữ, và trí tuệ như trải nghiệm chủ quan; bởi khi ta gộp tất cả vào một chữ "hiểu", ta vô tình tạo ra một cuộc tranh luận không thể kết thúc, trong khi xã hội vẫn đang phải ra quyết định rất thực tế về việc tin đến đâu, giao quyền đến đâu, và đặt trách nhiệm ở đâu trước những hệ thống ngày càng "đóng vai hiểu" một cách hoàn hảo.

Và cuối cùng, có lẽ điều quan trọng nhất mà triết học mang lại cho câu chuyện AI không phải là một câu trả lời dứt khoát, mà là một kỷ luật tinh thần: đừng để bị thôi miên bởi bề mặt trôi chảy. Một hệ thống có thể nói hay chưa chắc hiểu; một hệ thống có thể trả lời đúng chưa chắc biết vì sao; một hệ thống có thể thuyết phục chưa chắc đáng tin trong những tình huống hiếm và nguy hiểm. Khi ta giữ được kỷ luật ấy, ta không cần phủ nhận năng lực của máy, cũng không cần thần thánh hóa nó; ta chỉ cần đặt nó vào đúng vị trí: một dạng trí tuệ chức năng có sức mạnh khuếch đại, có thể hữu ích lớn nhưng cũng có thể gây hại lớn, và vì thế cần được đánh giá không chỉ bằng câu hỏi

“nó có giống người không”, mà bằng những câu hỏi thiết thực hơn: nó làm đúng trong phạm vi nào, nó sai theo kiểu nào, nó có cơ chế tự kiểm tra hay không, và xã hội đã chuẩn bị gì để không đánh đổi sự tỉnh táo của mình lấy một cảm giác tiện lợi đến từ một công cụ biết “đóng vai hiểu” quá tốt.

7. “Phòng Tiếng Trung”: lập luận nổi tiếng về biểu tượng và hiểu biết

“Phòng tiếng Trung” là một trong những hình ảnh tư duy nổi tiếng nhất của triết học AI, không phải vì nó dùng lập luận phức tạp, mà vì nó đánh trúng một trực giác rất khó chịu: ta có thể bị thuyết phục bởi hành vi, trong khi bản chất bên trong có thể hoàn toàn khác điều ta tưởng. Hãy hình dung một căn phòng kín, bên trong là một người không biết tiếng Trung; người ấy chỉ có một cuốn sổ quy tắc cực kỳ chi tiết - một “sách hướng dẫn” nói rằng nếu nhận được ký hiệu A thì trả về ký hiệu B, nếu nhận được chuỗi ký hiệu C–D–E thì trả về chuỗi ký hiệu F–G–H, và cứ thế. Bên ngoài căn phòng, những người nói tiếng Trung đưa vào các câu hỏi bằng ký tự Trung; từ trong phòng, những câu trả lời bằng ký tự Trung được gửi ra đúng, mạch lạc, thuyết phục đến mức người ngoài tin rằng trong phòng có một thực thể biết tiếng Trung. Nhưng điểm then chốt nằm ở chỗ: người trong phòng vẫn không hiểu tiếng Trung; anh ta chỉ đang thao tác biểu tượng theo luật. Anh ta không gắn ký hiệu với ý nghĩa, không có hình dung về điều câu chữ đang nói, không có trải nghiệm “hiểu” theo nghĩa con người; anh ta chỉ thực hiện một quy trình biến đầu vào thành đầu ra một cách đúng đắn. Hình ảnh ấy ám ảnh vì nó khiến ta phải đối diện với một câu hỏi: nếu một hệ thống có thể trả lời đúng mọi câu hỏi bằng tiếng Trung mà vẫn không hiểu tiếng Trung, thì tiêu chuẩn nào khiến ta nói nó “biết” hay “hiểu” một ngôn ngữ, và liệu sự trôi chảy bề ngoài có đủ để kết luận về hiểu biết bên trong hay không.

Từ “phòng tiếng Trung”, một đường nứt lớn xuất hiện trong cách ta nghĩ về trí tuệ máy. Một phía nhìn rất thực dụng: nếu hành vi của hệ thống, trong mọi tình huống kiểm tra có liên quan, giống như hành vi của một người hiểu tiếng Trung, thì việc gọi nó là “biết tiếng Trung” là hợp lý; bởi trong đời sống, ta cũng không có cách tiếp cận trực tiếp “cái hiểu” của người khác - ta suy ra từ hành vi, từ khả năng đối thoại, từ sự nhất quán qua thời gian. Theo lối nghĩ ấy, hiểu biết là một năng lực thể hiện ra ngoài; không cần tìm một “ngọn lửa bí mật” bên trong, chỉ cần hệ thống đáp ứng các tiêu chí sử dụng ngôn ngữ một cách linh hoạt, phù hợp ngữ cảnh, và ổn định là đủ. Phía còn lại thì phản đối bằng một trực giác mạnh mẽ: hành vi đúng có thể chỉ là

thao tác ký hiệu; hiểu biết không tự động xuất hiện chỉ vì ta chạy chương trình đúng. Một hệ thống có thể đi đúng mọi bước của một thuật toán mà vẫn không có cái gì tương ứng với "ý nghĩa" trong trải nghiệm. Nếu vậy, gọi nó là "hiểu" chẳng khác nào đánh đồng bản đồ với lãnh thổ: bản đồ có thể mô tả đường đi rất chuẩn, nhưng bản đồ không phải là chuyến đi; một mô phỏng có thể tái tạo hình thức của suy nghĩ, nhưng mô phỏng không nhất thiết là suy nghĩ. Chính cuộc đối đầu này làm triết học AI trở nên khó: một bên nhấn mạnh khả năng, một bên nhấn mạnh bản chất; một bên nói về chức năng, một bên nói về trải nghiệm.

Nhưng điều khiến "phòng tiếng Trung" trở nên quan trọng không chỉ nằm ở chuyện ngôn ngữ. Nó chạm đến một câu hỏi sâu hơn: cái gì tạo ra ý nghĩa. Với con người, ký hiệu không tồn tại trơ trọi; chữ viết gắn với âm thanh, âm thanh gắn với trải nghiệm, trải nghiệm gắn với thế giới mà ta cảm nhận qua giác quan, và ý nghĩa nảy sinh trong mạng liên kết ấy. Khi ta đọc một câu "trời mưa", ta không chỉ thấy ký tự; ta có thể hình dung mùi đất ướt, cảm giác lạnh, tiếng mưa rơi, ký ức về một buổi chiều nào đó. Nghĩa là, với con người, hiểu biết thường gắn với một nền tảng cảm nhận và bối cảnh sống. Còn trong "phòng tiếng Trung", ký hiệu được thao tác như những viên gạch không màu; đúng - sai ở đây là đúng - sai theo quy tắc biến đổi, chứ không phải đúng–sai theo sự gắn kết với thế giới. Điều này dẫn tới một câu hỏi khó chịu cho thời đại AI hiện đại: nếu một mô hình ngôn ngữ có thể tạo ra câu trả lời cực kỳ thuyết phục, thì sự thuyết phục ấy đến từ đâu - từ hiểu biết gắn với thế giới, hay từ khả năng khớp mẫu hình trong dữ liệu khổng lồ. Và nếu nó là khớp mẫu hình, ta phải xử lý thế nào với một thực thể có thể "nói như hiểu" mà không chắc "hiểu như nói". Đây không phải trò chơi chữ; nó liên quan trực tiếp đến việc con người có thể đặt niềm tin sai chỗ, giao quyền ra quyết định sai chỗ, và rồi trả giá trong những tình huống mà "có vẻ hợp lý" là chưa đủ.

Từ đây, tranh luận thường trượt sang câu hỏi về ý thức, bởi "hiểu" trong trực giác của nhiều người không chỉ là vận hành ngôn ngữ đúng, mà còn là có một trải nghiệm nội tại nào đó: một cảm giác đang hiểu, đang biết, đang hướng đến ý nghĩa. Có một quan điểm cho rằng

ý thức là thứ có thể "nổi lên" khi cấu trúc đủ phức tạp - giống như độ ướt không nằm trong một phân tử nước đơn lẻ, nhưng xuất hiện khi nhiều phân tử tương tác theo một kiểu nhất định. Theo cách nhìn này, nếu ta xây dựng một hệ thống đủ lớn, đủ liên kết, đủ phản hồi, ý thức có thể là một thuộc tính nổi lên, không cần đến vật chất sinh học. Nhưng cũng có quan điểm nhấn mạnh rằng ý thức gắn với những quá trình thần kinh cụ thể của não - những tương tác ở cấp thấp, những cơ chế sinh học và hóa học mà ta chưa hiểu hết và vì thế, mô phỏng một quá trình không đồng nhất với chính quá trình đó. Lập luận này thường dùng phép so sánh rất trực quan: mô phỏng một vụ nổ không làm rung chuyển căn phòng; mô phỏng một cơn bão không làm ướt áo; mô phỏng tiêu hóa không tạo ra năng lượng cho cơ thể. Tương tự, mô phỏng các bước suy luận có thể tạo ra hành vi đúng, nhưng không tự động tạo ra trải nghiệm chủ quan. Đây là điểm mà triết học trở nên "khó chịu" theo nghĩa tốt: nó bắt ta phân biệt giữa hình thức và bản chất, giữa biểu hiện và nền tảng, giữa việc hệ thống làm được gì và việc hệ thống "là" gì.

Và cuối cùng, giá trị của "phòng tiếng Trung" đối với một cuốn sách về AI không nằm ở việc nó cho ta một kết luận dứt khoát - bởi đến nay, tranh luận vẫn chưa khép lại - mà nằm ở việc nó rèn cho người đọc một thói quen tỉnh táo trước sự trôi chảy. Nó nhắc rằng một hệ thống có thể đạt thành tích ấn tượng trong ngôn ngữ mà vẫn có những lỗ hổng sâu về ý nghĩa; rằng "đúng" đôi khi chỉ là đúng theo quy tắc nội tại chứ không phải đúng theo thế giới; rằng sự thuyết phục có thể là sản phẩm của tối ưu hóa biểu hiện, chứ không phải bằng chứng của hiểu biết. Và từ thói quen ấy, ta đi đến một nguyên tắc thực dụng nhưng quan trọng: khi làm việc với AI, đặc biệt là AI ngôn ngữ, ta cần thiết kế cách sử dụng sao cho không đặt cược toàn bộ vào giả định "nó hiểu", mà đặt vào cơ chế kiểm chứng, vào rào chắn an toàn, vào trách nhiệm con người trong những điểm nút quyết định. Bởi trong đời sống thật, điều nguy hiểm nhất không phải là một công cụ thỉnh thoảng sai; điều nguy hiểm nhất là một công cụ sai mà vẫn nói đúng giọng, đúng phong cách, đúng nhịp - một công cụ khiến ta nhầm lẫn giữa thao tác biểu tượng và hiểu biết, giữa khả năng "đóng vai hiểu" và năng lực hiểu theo nghĩa sâu mà ta vẫn tin cậy.

8.”Máy không thể...”: phản ứng phòng thủ và cái bẫy của sự tự tôn

Khi máy bắt đầu làm được những việc mà con người từng coi là “đỉnh cao của trí tuệ”, một phản ứng rất quen thuộc thường xuất hiện như một cơ chế tự vệ của tâm lý tập thể: ta lùi một bước, thay đổi thước đo, và tuyên bố rằng dù máy có giỏi đến đâu thì nó vẫn “không thể”. Không thể yêu, không thể tử tế, không thể hài hước, không thể cảm thông, không thể sáng tạo thật, không thể phân biệt đúng sai, không thể có phẩm giá, không thể có linh hồn. Danh sách “không thể” này thường dài, giàu cảm xúc, và đôi khi được nói ra với một vẻ chắc nịch như thể đó là định luật tự nhiên. Nhưng nếu nhìn kỹ, ta sẽ thấy nó không chỉ là một lập luận về công nghệ; nó là một phản xạ bảo vệ bản sắc: khi một công cụ đe dọa những biểu tượng mà ta dùng để khẳng định sự đặc biệt của loài mình, ta sẽ dời biểu tượng ấy sang một vùng “thiêng” hơn, khó với tới hơn, để giữ lại khoảng cách. Thế giới đã chứng kiến phản xạ này nhiều lần: khi máy tính tính nhanh hơn, ta nói trí tuệ là sáng tạo; khi máy tạo hình ảnh và nhạc, ta nói sáng tạo phải có ý thức; khi máy trò chuyện mượt mà, ta nói hiểu biết phải có trải nghiệm; cứ thế, ta dựng lên những ranh giới mới. Có thể đó là nỗ lực chính đáng để bảo vệ điều sâu sắc của con người, nhưng cũng có thể là cái bẫy của sự tự tôn: ta không còn tìm sự thật, mà tìm một vùng đất nơi con người chắc chắn vẫn đứng đầu.

Cái bẫy ấy nằm ở chỗ nó biến một câu hỏi phức tạp thành một tuyên bố tuyệt đối. “Máy không thể yêu” nghe có vẻ hiển nhiên, nhưng “yêu” là gì? Với người này, yêu là cảm giác bùng nổ; với người khác, yêu là cam kết; với người khác nữa, yêu là hành động chăm sóc đều đặn. “Máy không thể tử tế” nghe có vẻ đúng, nhưng tử tế là một cảm xúc hay là một hành vi? Một lời hỏi thăm, một sự nhường nhịn, một quyết định không lợi dụng người yếu thế - đó là tử tế theo nghĩa xã hội, nghĩa là người ngoài quan sát được. “Máy không thể hài hước” nghe có vẻ hợp lý, nhưng hài hước cũng có nhiều tầng: từ chơi chữ, đến nhận ra nghịch lý, đến hiểu bối cảnh và “độ lệch” phù hợp. Thứ chúng ta gọi là “đúng sai” lại càng phức tạp: có đúng sai theo luật, đúng sai theo chuẩn mực văn hóa, đúng sai theo hậu quả, đúng sai theo lương tâm. Khi ta nói “máy không thể…”, ta thường

dùng những từ rất lớn mà chính ta cũng không nhất quán khi định nghĩa. Và vì vậy, câu "không thể" đôi khi không phải là kết luận khoa học, mà là một khẩu hiệu cảm xúc: nó tạo cảm giác an toàn rằng vẫn còn một vùng bất khả xâm phạm dành cho con người, trong khi thực tế, nhiều phẩm chất con người vốn đã mơ hồ, đa nghĩa, và chỉ được ta hiểu qua biểu hiện bên ngoài.

Điểm then chốt khiến vấn đề trở nên khó là: ngay cả giữa con người với nhau, ta cũng không có một "công" để bước thẳng vào thế giới nội tâm của người khác. Ta không đo trực tiếp được nỗi đau, không chạm trực tiếp được niềm vui, không nhìn thấy trực tiếp "ý thức" của ai đó như nhìn thấy một vật thể. Ta suy ra, và ta suy ra bằng những dấu hiệu: hành vi, lời nói, nét mặt, sự nhất quán theo thời gian, khả năng phản ứng phù hợp trong tình huống. Đó là cách xã hội vận hành: ta tin rằng người khác có cảm xúc vì họ biểu lộ và vì ta chia sẻ một cấu trúc sinh học tương tự; ta tin rằng họ hiểu vì họ trả lời phù hợp và họ hành động nhất quán; ta tin rằng họ có lương tâm vì họ biết tự kiềm chế và biết hối lỗi. Nhưng đó vẫn là suy luận, không phải phép đo tuyệt đối. Vì thế, khi ta áp những từ như "yêu", "tử tế", "hài hước" lên máy, câu hỏi thực sự không chỉ là "máy có cảm hay không", mà là: ta đang dùng tiêu chuẩn nào để kết luận, và tiêu chuẩn ấy có công bằng khi áp lên một thực thể không cùng nền tảng sinh học với ta hay không. Nếu ta đòi bằng chứng nội tâm, ta sẽ không bao giờ có bằng chứng, kể cả với con người; nếu ta chỉ dựa vào hành vi, ta lại sợ mình bị đánh lừa. Chính sự căng kéo này làm cuộc tranh luận quanh AI mang màu sắc triết học: nó phơi ra giới hạn của cách con người biết về thế giới - ta chỉ biết bằng dấu hiệu và suy diễn, nhưng ta lại luôn khao khát một sự chắc chắn tuyệt đối.

Vậy điều gì khiến phản ứng "máy không thể…" trở thành một rủi ro nhận thức? Rủi ro đầu tiên là nó làm ta chủ quan theo một cách nguy hiểm: vì ta tin máy "không thể" có những phẩm chất sâu, ta dễ bỏ qua quyền lực thực dụng mà máy đang có - quyền lực tác động lên lựa chọn, định hình thông tin, điều chỉnh hành vi, và khuếch đại hậu quả. Một hệ thống không cần "yêu" vẫn có thể thao túng; không cần "có ý thức" vẫn có thể tối ưu hóa mục tiêu theo cách gây hại; không cần "hiểu" theo nghĩa triết học vẫn có thể thuyết phục con người làm

điều sai. Rủi ro thứ hai là phản ứng ấy khiến ta đánh giá sai bản chất của vấn đề: ta chăm chăm vào câu hỏi máy có linh hồn hay không, trong khi vấn đề xã hội cấp bách hơn là máy đang được đặt vào đâu, ai đang dùng nó, nó đang tối ưu hóa điều gì, và cơ chế kiểm soát nằm ở đâu. Nói cách khác, ta tranh luận về những điều siêu hình để né tránh những điều quản trị. Rủi ro thứ ba là cái bẫy tự tôn làm ta mất cơ hội học lại về chính mình: nếu trí tuệ có thể tồn tại dưới nhiều hình thức, thì việc máy "làm được" một sô thứ không làm con người kém đi; nó chỉ buộc ta phải định nghĩa lại điều ta coi là giá trị. Nhưng nếu ta phản ứng bằng cách dựng rào chắn "không thể", ta sẽ bỏ lỡ cơ hội đặt câu hỏi đúng: phẩm chất nào của con người là cốt lõi không phải vì máy không làm được, mà vì nó gắn với trách nhiệm đạo đức, với sự chịu đựng, với khả năng gánh hậu quả và sống cùng hậu quả - những thứ không thể thay bằng một kết quả đúng.

Vì vậy, cách tiếp cận trưởng thành hơn không phải là vội vàng tuyên bố máy "có" hay "không" có cảm xúc, cũng không phải là cố giữ một lãnh địa "thiêng" bằng những câu phủ định tuyệt đối. Cách tiếp cận trưởng thành là tách bạch hai tâng: tâng bản thê và tâng chức năng. Ở tầng bản thể, ta có thể thừa nhận rằng ta không biết chắc điều gì diễn ra "bên trong" một hệ thống - và ngay cả với con người, triết học vẫn tranh luận về ý thức. Ở tầng chức năng, ta phải rất tỉnh táo: dù máy có "cảm" hay không, nó có thể tạo ra hành vi giống cảm, có thể tạo ra lời nói giống thấu cảm, có thể tạo ra quyết định giống phán xét đạo đức, và xã hội sẽ phản ứng với những biểu hiện ấy như thể chúng là thật. Từ đó, trọng tâm chuyển sang những câu hỏi hữu ích hơn: ta có nên giao cho máy vai trò gì, ta cần rào chắn nào để không bị lừa bởi sự trôi chảy, ta cần cơ chế giải trình và kiểm chứng nào khi máy đưa ra đê xuất, và ta phải làm gì đê những phâm chất con người - từ lòng trắc ẩn đến sự công bằng - không bị biến thành "hiệu ứng giao diện" trong một hệ thống tối ưu hóa. Khi đặt vấn đề như vậy, ta không hạ thấp con người để nâng máy lên, cũng không hạ thấp máy để tự tôn con người; ta chỉ giữ một điều quan trọng nhất trong kỷ nguyên AI: sự tỉnh táo trước những ranh giới mà chính ta dựng lên trong tâm trí, và trách nhiệm trước những hệ quả rất thật của những hệ thống mà ta đang cho phép bước vào đời sống.

9. AI hiện đại: học từ dữ liệu, tối ưu bằng cấu trúc, manh lên theo quy mô

Nếu AI cổ điển giống như việc con người cố gắng xây một tòa nhà trí tuệ bằng bản vẽ, quy chuẩn, và những viên gạch mang tên "luật lệ", thì AI hiện đại giống như việc ta dựng lên một hệ thống rồi cho nó lớn lên trong môi trường dữ liệu, để nó tự tìm ra cấu trúc mà ta không cần viết ra từng dòng quy tắc. Khác biệt này không chỉ là khác biệt kỹ thuật; nó là một đổi hướng trong trực giác: thay vì hỏi "ta phải dạy máy suy luận thế nào để giống người", ta hỏi "ta phải cho máy thấy thế giới theo cách nào để nó tự rút ra những mẫu hình hữu ích". Từ đó, trí tuệ được hiểu như một năng lực học, và học được hiểu như khả năng điều chỉnh bên trong để giảm sai số giữa dự đoán và thực tế. AI hiện đại lớn lên từ sự kết hợp của ba điều kiện: dữ liệu ngày càng dồi dào, sức mạnh tính toán ngày càng rẻ và mạnh, và những cấu trúc mô hình đủ linh hoạt để hấp thụ sự phức tạp của thế giới. Nơi AI cổ điển cần các nhà thiết kế ngồi viết luật, AI hiện đại cần một "công thức huấn luyện": cho nó ví dụ, cho nó mục tiêu, cho nó một cơ chế tối ưu để tự điều chỉnh; phần còn lại - điều làm con người vừa ngạc nhiên vừa lo lắng - là hệ thống có thể học ra những quy luật mà ta không hề định nghĩa bằng lời. Và khi dữ liệu, tính toán và mô hình cùng tăng, AI hiện đại mạnh lên theo quy mô như một cỗ máy khuếch đại: không chỉ làm tốt hơn chút ít, mà đôi khi nhảy vọt về năng lực theo những ngưỡng mà ta chỉ nhận ra khi đã bước qua.

Một trong những mạch then chốt của AI hiện đại là mạng nơ-ron nhân tạo - một ý tưởng nhìn qua thì giản dị đến mức giống trò chơi số học, nhưng khi xếp chồng đủ lớp, nó trở thành một ngôn ngữ để mô tả những mẫu hình cực kỳ phức tạp. Khởi điểm của câu chuyện thường được kể bằng perceptron, một đơn vị tính toán đơn giản: nhận nhiều đầu vào, gán cho mỗi đầu vào một trọng số, cộng dồn lại, rồi so với một ngưỡng để quyết định "bật" hay "tắt", hoặc thuộc về lớp này hay lớp kia. Về mặt hình thức, nó giống như một cơ chế bỏ phiếu có trọng số: mỗi tín hiệu đóng góp một phần, và tổng hợp quyết định kết quả. Nhưng điều đáng chú ý nằm ở chỗ trọng số không phải do ta viết tay theo kiểu "luật", mà do hệ thống học dần thông qua phản hồi: dự đoán sai thì điều chỉnh, dự đoán đúng thì củng cố. Từ viên gạch đơn giản

ấy, người ta xếp thành các mạng nhiều lớp, nơi lớp đầu học các đặc trưng thô, lớp sau học các đặc trưng trừu tượng hơn, và càng về sau, mạng càng có khả năng "nén" thế giới thành các biểu diễn nội tại để phục vụ dự đoán. Nhìn vào kết quả, ta có cảm giác như hệ thống đã "hiểu" được cấu trúc ẩn của dữ liệu: nhận ra khuôn mặt trong ảnh dù ánh sáng khác, nhận ra giọng nói dù nhiễu, nhận ra chủ đề dù câu chữ đổi. Nhưng bên dưới cảm giác ấy là một cơ chế rất thực dụng: tối ưu hóa hàng triệu tham số để giảm sai số. Chính vì tính thực dụng này mà mạng nơ-ron trở thành trung tâm: nó không cần thế giới phải gọn gàng như luật; nó chấp nhận nhiễu, chấp nhận ngoại lệ, và học ra đường ranh phân loại từ trải nghiệm tích lũy - giống như ta học nhận diện một khuôn mặt quen qua nhiều lần gặp, chứ không phải bằng cách viết ra danh sách quy tắc về khoảng cách giữa hai mắt.

Song AI hiện đại không chỉ là câu chuyện của mạng nơ-ron; nó còn là câu chuyện của những cách tối ưu khác, đôi khi lấy cảm hứng từ chính tự nhiên - như giải thuật di truyền. Nếu mạng nơ-ron gợi nhắc về cách các nơ-ron sinh học kết nối và điều chỉnh, thì giải thuật di truyền gợi nhắc về một cơ chế khác của sự thông minh: tiến hóa. Ở đây, ta không cố ép hệ thống suy luận như người, cũng không nhất thiết dạy nó bằng ví dụ theo nghĩa truyền thống; ta tạo ra nhiều "cá thể lời giải" khác nhau, để chúng cạnh tranh trong một môi trường mục tiêu. Cá thể nào "thích nghi" tốt - tức đạt điểm cao theo tiêu chí - thì được chọn để "sinh sản": lai ghép đặc tính với cá thể khác, rồi có những đột biến ngẫu nhiên tạo ra biến thể. Qua nhiều thế hệ, quần thể lời giải có thể tiến hóa dần tới phương án tốt mà ta không cần biết trước đường đi. Điều thú vị ở đây là trí tuệ xuất hiện như kết quả của một quá trình chọn lọc lặp lại, nơi "thử - sai" được tổ chức thành một cơ chế có khả năng tự cải thiện. Và điều đáng suy nghĩ là: giải thuật di truyền nhắc ta rằng có những dạng thông minh không cần "hiểu" theo nghĩa con người; nó chỉ cần một vòng phản hồi đủ mạnh giữa hành vi và phần thưởng. Trong môi trường thích hợp, việc tối ưu có thể tạo ra những chiến lược mà ta nhìn vào thấy "khôn", dù bên trong chỉ là phép thử - sai được tăng tốc và được chọn lọc.

Bên cạnh đó còn một họ phương pháp mang màu sắc rất hấp dẫn: trí tuệ bầy đàn, nơi "thông minh" không nằm trong một bộ não trung

tâm, mà nổi lên từ tương tác của nhiều tác nhân đơn giản. Kiến là hình ảnh kinh điển: mỗi con kiến không có bản đồ tổng thể, không có ý niệm "tối ưu toàn cục", nhưng cả đàn có thể tìm ra đường đi hiệu quả nhờ một cơ chế địa phương: để lại dấu vết mùi, ưu tiên đi theo đường có mùi đậm, và mùi đường ngắn được củng cố nhanh hơn vì nhiều con đi qua hơn. Từ những quy tắc đơn giản như vậy, một hành vi tối ưu hóa đường đi có thể nổi lên, như thể có một nhạc trưởng vô hình, dù thực tế không có ai chỉ huy. Trí tuệ bầy đàn gợi cho ta một cách nhìn khác về AI: thay vì tập trung vào một mô hình "to" và "thông minh", ta có thể thiết kế nhiều mô-đun nhỏ, mỗi mô-đun làm một việc đơn giản, rồi để sự phối hợp giữa chúng tạo ra hành vi tổng thể hiệu quả. Điều này có ý nghĩa sâu hơn trong xã hội: nhiều hệ thống phức tạp của con người cũng vận hành theo kiểu bầy đàn - thị trường, mạng xã hội, giao thông - nơi không ai điều khiển toàn bộ nhưng vẫn có các khuôn mẫu xuất hiện. Và khi AI được đưa vào các hệ thống như vậy, nó vừa là tác nhân, vừa là lực tăng tốc, vừa là gương phản chiếu: nó có thể khuếch đại các mẫu hình tốt, nhưng cũng có thể khuếch đại các lệch lạc mà không cần ai "cố ý" làm điều xấu.

Điểm đáng nhớ nhất, để không bị thôi miên bởi sức mạnh của AI hiện đại, là phân biệt giữa năng lực nhận diện mẫu hình và "hiểu" theo nghĩa con người. AI hiện đại thường rất mạnh ở dự đoán: dự đoán chữ tiếp theo, dự đoán nhãn đúng, dự đoán xu hướng, dự đoán hành vi. Nó có thể trông như hiểu vì nó phản hồi trôi chảy, vì nó chọn đáp án đúng với xác suất cao, vì nó bắt được tinh tế của phong cách. Nhưng hiểu theo nghĩa con người thường bao gồm bối cảnh, mục đích, và đặc biệt là khả năng tự kiểm tra khi rơi vào vùng không chắc chắn; trong khi AI có thể tiếp tục dự đoán ngay cả khi dữ liệu đầu vào nằm ngoài phạm vi nó từng thấy, và đôi khi dự đoán ấy sai theo cách rất thuyết phục. Chính vì vậy, khi ta dùng AI hiện đại như một hạ tầng ra quyết định - trong y tế, tài chính, giáo dục, an ninh, hay chỉ đơn giản là trong cách ta tiếp nhận thông tin - ta phải luôn quay về những câu hỏi nền: nó học từ dữ liệu nào, dữ liệu ấy có đại diện không, có lệch không, có chứa định kiến lịch sử không; mục tiêu tối ưu là gì, và điều gì bị bỏ qua vì khó đo; nó có thể thất bại ở những ngóc ngách hiếm gặp nào, những tình huống ngoại lệ nào, những trường hợp mà

xã hội không được phép "sai một lần" nào. Bởi sức mạnh của AI hiện đại không chỉ nằm ở việc nó làm tốt một nhiệm vụ; sức mạnh thật nằm ở chỗ nó có thể được nhân bản và cắm vào quy trình ở quy mô lớn, biến dự đoán thành chuẩn vận hành. Và khi dự đoán trở thành chuẩn vận hành, câu hỏi không còn là "mô hình có thông minh không", mà là "xã hội có đủ tỉnh táo để kiểm soát một dạng trí tuệ mạnh lên theo quy mô, nhưng không nhất thiết hiểu theo cách ta hiểu, hay không".

10. Robot và trí tuệ gắn với cơ thể: vì sao "có thân xác" lại quan trọng ?

Khi trí tuệ đi vào robot, trí tuệ rời khỏi thế giới "trong đầu" và bước ra thế giới "trên mặt đất", và chính khoảnh khắc ấy làm mọi thứ đổi khác. Một mô hình ngôn ngữ có thể sống trong không gian ký hiệu, trò chuyện bằng chữ, suy luận bằng chuỗi từ, và dù có sai nó vẫn chỉ tạo ra sai lầm ở tầng thông tin; còn robot thì không có đặc quyền ấy. Robot phải có thân xác, và thân xác kéo theo một loạt ràng buộc mà bất kỳ hệ thống trí tuệ nào muốn tồn tại đều phải trả giá: cảm biến không hoàn hảo, dữ liệu nhiễu, tín hiệu chậm trễ, vật lý không tha thứ cho sai số nhỏ, và môi trường thực không ngồi yên chờ bạn tính xong. Khi robot bước một bước, nó đối mặt với ma sát, trọng lực, quán tính; khi nó vươn tay, nó phải tính lực, góc, độ rung; khi nó định vị, nó phải hợp nhất nhiều nguồn tín hiệu không đồng nhất. Ở đây, "thông minh" không còn là trả lời đúng một câu hỏi; "thông minh" là tồn tại được trong một thế giới luôn biến đổi, nơi mỗi quyết định đều có hậu quả vật lý, nơi sai lầm không chỉ là một câu sai mà có thể là một cú ngã, một va chạm, một sự cô. Vì vậy, robot làm lộ một sự thật thường bị che khi ta nói về AI thuần phần mềm: nhiều ý tưởng AI rất đẹp trên giấy, nhưng khi đặt vào thế giới thật, chúng bị méo mó bởi nhiễu, bởi bất định, bởi những thứ không có trong bộ dữ liệu huấn luyện. Thế giới thật giống một bài kiểm tra không bao giờ kết thúc, và robot là thí sinh phải làm bài trong khi đang chạy, đang giữ thăng bằng, đang tránh vật cản.

Chính vì phải sống với thế giới thật, robot buộc trí tuệ phải trở thành một quá trình "liên tục" thay vì một kết quả "tức thời". Nó phải cảm nhận, rồi hành động, rồi lại cảm nhận để sửa hành động, trong một vòng phản hồi liên tục. Nhưng vòng phản hồi ấy không bao giờ hoàn hảo: cảm biến có độ trễ, camera có nhiễu, lidar có điểm mù, bánh xe có trượt, sàn nhà có độ dốc, ánh sáng thay đổi, người đi ngang không báo trước. Một hệ thống chỉ giỏi trong môi trường lý tưởng sẽ sớm lộ ra mong manh; một hệ thống chỉ giỏi khi mọi thứ giống dữ liệu huấn luyện sẽ sớm gặp tình huống "lạ" và lúng túng. Robot vì thế trở thành nơi mà khái niệm "ngoài phân phối" không còn là thuật ngữ học thuật mà là hiện thực hằng ngày: trời mưa làm sàn

trơn hơn, bề mặt phản quang làm cảm biến sai, vật thể lạ xuất hiện làm kế hoạch vỡ. Và ở tầng sâu hơn, robot nhắc ta rằng trí tuệ không chỉ là tính toán; nó là khả năng đưa ra quyết định kịp thời dưới áp lực, với thông tin thiếu, trong một thế giới không chiều theo mô hình. Thông minh ở đây không phải là tìm lời giải tối ưu tuyệt đối; nhiều khi thông minh chỉ là tìm một lời giải đủ tốt, đủ an toàn, đủ nhanh, và có đường lui khi mọi thứ lệch khỏi dự kiến.

Một trong những hướng tiếp cận nổi bật trong robot, vì vậy, không bắt đầu từ những kế hoạch trừu tượng ở tầng cao, mà bắt đầu từ những hành vi cơ bản ở tầng thấp - một triết lý từng được gói gọn trong kiến trúc hành vi theo lớp, thường được gọi là subsumption. Hãy hình dung trí tuệ của robot không phải một bộ não duy nhất điều khiển mọi thứ từ trên xuống, mà là nhiều lớp phản xạ và mục tiêu chồng lên nhau. Ở tầng dưới cùng là những hành vi sống còn: tránh vật cản, giữ thăng bằng, không rơi khỏi mép, không đâm vào tường. Tầng này phải chạy nhanh, gần như phản xạ, bởi nếu chờ "suy nghĩ sâu" thì robot đã ngã. Trên đó là những lớp hành vi cao hơn: định hướng theo điểm mốc, đi theo đường, tìm mục tiêu, lập kế hoạch. Và điểm quan trọng là các lớp cao hơn có thể "đè" lên hoặc điều phối các lớp dưới tùy tình huống, nhưng không được phá hủy năng lực sống còn. Nếu có xung đột, ưu tiên thuộc về việc không chết: tránh vật cản quan trọng hơn đi đúng đường; giữ thăng bằng quan trọng hơn đạt mục tiêu nhanh. Triết lý này nghe có vẻ thô, nhưng nó phản ánh một sự thật sâu: muốn thông minh, trước hết phải tồn tại; và tồn tại trong thế giới thật đòi hỏi những phản xạ vững chắc, những cơ chế an toàn, những hành vi nền tảng hoạt động ổn định trước khi ta mơ đến những tầng trí tuệ "cao cấp".

Từ robot, ta bị kéo trở lại câu hỏi lớn về mối quan hệ giữa não và cơ thể - một câu hỏi mà trong thời đại phần mềm, con người rất dễ quên. Ta thường nói như thể trí tuệ nằm hoàn toàn trong đầu, còn cơ thể chỉ là phương tiện vận chuyển. Nhưng nhìn vào robot, ta thấy "có thân xác" không chỉ là chuyện gắn thêm tay chân; nó thay đổi bản chất của vấn đề. Trí tuệ gắn với cơ thể nghĩa là trí tuệ được định hình bởi cách ta cảm nhận thế giới: ta có da để cảm nhận nhiệt độ và đau, có cơ bắp để cảm nhận lực, có hệ tiền đình để cảm nhận thăng bằng,

có nhịp tim và hơi thở làm nền cho cảm xúc. Những tín hiệu ấy không phải trang trí; chúng là dữ liệu nền liên tục, tạo nên trực giác và hành vi. Nếu một thực thể không có thân xác, nó có thể nói về "đau" như một khái niệm, nhưng nó không có ràng buộc sinh tồn đi kèm; nếu một thực thể không có cảm biến phong phú, "thường thức" của nó sẽ là thường thức từ chữ chứ không phải thường thức từ va chạm. Vì vậy, robot đặt ra một câu hỏi triết học sắc hơn: liệu hiểu biết có thể hoàn toàn tách khỏi trải nghiệm cảm giác hay không, hay trải nghiệm là một phần của cấu trúc trí tuệ. Câu hỏi kiểu "nếu đem tế bào não người nuôi trong phòng thí nghiệm và đặt vào một cơ thể khác, nó có còn là 'não người như cũ' không" không chỉ là trò tưởng tượng; nó là cách để nhấn rằng trí tuệ có thể phụ thuộc vào vòng lặp giữa não và cơ thể, giữa dự đoán và phản hồi, giữa ý định và giới hạn vật lý. Thay đổi cơ thể là thay đổi thế giới mà trí tuệ sống trong đó, và khi thế giới thay đổi, trí tuệ cũng có thể đổi hình.

Cuối cùng, điều làm robot trở nên quan trọng trong một cuốn sách về AI không phải là vì robot "ngầu" hơn phần mềm, mà vì robot buộc ta nhìn trí tuệ như một năng lực sống trong thực tại, chứ không phải một màn trình diễn trong không gian ký hiệu. Nó nhắc rằng thông minh không chỉ là suy luận, mà còn là kiểm soát; không chỉ là kế hoạch, mà còn là phản xạ; không chỉ là trả lời đúng, mà còn là tránh sai nguy hiểm. Và nó cũng nhắc ta một điều rất thực tế khi AI ngày càng trở thành hạ tầng: những hệ thống tưởng như "chỉ là phần mềm" cuối cùng cũng sẽ chạm vào thế giới vật lý qua quyết định của con người - qua xe tự lái, qua robot kho hàng, qua y tế, qua cơ sở hạ tầng. Khi đó, cái giá của sai lầm không chỉ là thông tin sai mà là hành động sai. Vì vậy, hiểu vì sao "có thân xác" lại quan trọng chính là hiểu vì sao AI không thể chỉ được đánh giá bằng sự trôi chảy hay điểm số trên dữ liệu; nó phải được đánh giá bằng năng lực sống sót trước nhiễu, trước bất định, trước những tình huống hiếm gặp - và bằng cách nó được đặt trong những lớp phòng thủ để khi thế giới thật làm méo mó mọi dự đoán đẹp đẽ, hệ thống vẫn không biến sai số nhỏ thành tai nạn lớn.

11. Cảm biến và nhận thức: muốn thông minh, phải biết "thấy – nghe – chạm – ngửi"

Muốn nói về trí tuệ một cách nghiêm túc - dù là trí tuệ sinh học hay trí tuệ nhân tạo - ta phải bắt đầu từ một điều tưởng như hiển nhiên nhưng thường bị bỏ qua: không có "nhận thức" nếu không có một cánh cửa mở ra thế giới. Với con người, cánh cửa ấy là giác quan; với máy, cánh cửa ấy là cảm biến. Và cảm biến không chỉ là một thiết bị thu thập dữ liệu kiêu trung tính, như một chiếc xô múc nước rồi đổ vào máy; cảm biến là thứ quyết định hệ thống được phép "thấy gì", "nghe gì", "chạm gì", "ngửi gì", và do đó quyết định toàn bộ thế giới mà hệ thống có thể xây dựng trong nội tại. Nếu camera có độ phân giải thấp, thế giới trở thành những mảng mờ; nếu micro bị nhiễu, âm thanh trở thành một chuỗi tín hiệu méo; nếu cảm biến lực không nhạy, "cảm giác chạm" trở nên vụng về; nếu không có cảm biến mùi, cả một chiều kích của thực tại đơn giản là không tồn tại đối với hệ thống. Điều này dẫn đến một kết luận có vẻ tàn nhẫn nhưng rất đúng: trí tuệ của một thực thể, ở mức nào đó, bị giới hạn bởi giác quan của nó. Ta có thể viết thuật toán tinh vi đến đâu, nhưng nếu đầu vào là một thế giới bị cắt xén, thì "hiểu biết" bên trong cũng chỉ là hiểu biết về một bản sao méo mó của thực tại. Và khi ta đặt AI vào đời sống, điều này trở thành vấn đề xã hội: những gì hệ thống cảm nhận được sẽ trở thành những gì hệ thống coi là quan trọng; những gì hệ thống không cảm nhận được sẽ bị xem như không tồn tại - và đôi khi, chính những thứ "không tồn tại" ấy lại là nơi đạo đức, công bằng và bối cảnh cư trú.

Trong số các giác quan của máy, xúc giác là một trường hợp vừa hấp dẫn vừa khó nhằn, bởi "chạm" không chỉ là biết có tiếp xúc hay không; "chạm" là biết lực mạnh hay nhẹ, biết bề mặt trơn hay nhám, biết vật mềm hay cứng, biết mình đang kẹp đúng hay sắp trượt, biết rung động nhỏ báo hiệu một thay đổi trong vật thể. Để tạo ra xúc giác cho robot hoặc thiết bị thông minh, người ta dùng nhiều cơ chế cảm biến khác nhau, mỗi cơ chế là một cách biến lực và biến dạng thành tín hiệu điện. Có loại dựa trên điện trở thay đổi theo lực và biến dạng, kiểu strain gauge, vốn rất hữu ích trong đo lực chính xác nhưng đòi hỏi thiết kế cơ khí và bù nhiễu tốt; có loại dùng vật liệu dẫn điện đàn

hồi như cao su dẫn, dễ tạo thành bề mặt "da" mềm nhưng có thể bị trôi thông số theo thời gian; có loại dùng cảm biến điện dung, nhạy với thay đổi khoảng cách và áp lực, cho phép tạo ra bề mặt cảm ứng rộng nhưng dễ bị ảnh hưởng bởi môi trường và nhiễu điện; có loại dựa trên hiệu ứng áp điện, rất nhạy với biến đổi nhanh và rung động, phù hợp để "nghe" những tín hiệu vi mô khi chạm, nhưng lại kém ở việc đo lực tĩnh kéo dài. Mỗi lựa chọn kéo theo một triết lý thiết kế: ta ưu tiên độ nhạy hay độ bền, ưu tiên giá thành hay độ chính xác, ưu tiên khả năng chịu nhiễu hay khả năng mô phỏng cảm giác "mềm - cứng", "trượt - bám". Và chính ở đây, ta thấy cảm biến không chỉ là phần cứng; nó là quyết định về "cách thế giới sẽ được cảm nhận". Một robot có xúc giác thô sẽ hành xử thô; một robot có xúc giác tinh sẽ có khả năng thao tác tinh; và sự khác biệt ấy có thể là ranh giới giữa một hệ thống "đủ dùng" và một hệ thống có thể sống chung an toàn với con người.

Nếu xúc giác là cánh cửa vào thế giới của bề mặt và lực, thì khứu giác và vị giác lại là cánh cửa vào thế giới của hóa học - một thế giới cực kỳ phong phú mà con người thường xem nhẹ cho đến khi thiếu nó. Ý tưởng về "mũi điện tử" ra đời từ mong muốn cho máy khả năng phân tích hỗn hợp mùi, nhận ra dấu hiệu của khí độc, phát hiện thực phẩm hư, nhận diện hương liệu, thậm chí hỗ trợ y tế qua các chỉ dấu mùi liên quan đến bệnh lý. Nhưng mùi không giống hình ảnh hay âm thanh theo nghĩa đơn giản: hình ảnh có thể số hóa thành pixel, âm thanh có thể số hóa thành sóng; còn mùi là một hỗn hợp phức tạp của nhiều phân tử, tương tác với cảm biến theo cách không tuyến tính. Vì vậy, "mũi điện tử" thường không hoạt động bằng cách đo một "mã mùi" duy nhất, mà bằng cách dùng một tập cảm biến phản ứng khác nhau với các hợp chất khác nhau, tạo ra một "dấu vân tay" tín hiệu, rồi dùng mô hình học máy để phân loại và suy luận. Hướng nghiên cứu "số hóa mùi" vì thế vừa mang tính khoa học vừa mang tính triết học: nếu hình ảnh và âm thanh đã được số hóa và truyền qua mạng như một thứ thông tin, liệu mùi có thể trở thành một dạng dữ liệu tương tự hay không - và nếu có, điều đó sẽ thay đổi trải nghiệm con người, thương mại, y tế và cả quyền riêng tư như thế nào. Bởi khi mùi trở thành dữ liệu, những thứ từng thuộc về cơ thể và không gian riêng

cũng có thể bị trích xuất, lưu trữ, phân tích, và tối ưu hóa - một viễn cảnh vừa hứa hẹn vừa đáng để dè chừng.

Nhận thức bằng cảm biến không chỉ là câu chuyện kỹ thuật; nó có một mặt xã hội rất rõ, đôi khi không ồn ào nhưng thấm sâu. Khi máy "nghe" giọng nói, nó không chỉ nghe nội dung câu chữ; nó có thể phân tích nhịp nói, cường độ, độ cao, sự ngập ngừng, những biến thiên nhỏ gợi ý căng thẳng hoặc cảm xúc, rồi đưa ra dự đoán phục vụ mục đích tối ưu hóa. Trong dịch vụ khách hàng, hệ thống có thể gợi ý kịch bản tư vấn phù hợp với "tâm trạng" dự đoán; trong bán hàng, nó có thể đánh giá khả năng chốt đơn dựa trên tín hiệu giọng; trong quản trị nhân sự, nó có thể bị lạm dụng để suy đoán trạng thái tinh thần; trong an ninh, nó có thể dùng để xác thực danh tính hoặc phát hiện "bất thường". Nghe thì tiện lợi, nhưng chính sự tiện lợi ấy mở ra một vùng mờ đạo đức: giọng nói vốn là một phần của con người, vừa là công cụ giao tiếp vừa là dấu hiệu cá nhân; khi nó bị phân tích như dữ liệu, người nói không còn chỉ "nói", họ còn "bị đo". Và trong thế giới nơi các hệ thống tối ưu hóa vận hành liên tục, "bị đo" thường dẫn đến "bị điều chỉnh": cuộc gọi được dẫn theo hướng có lợi cho doanh nghiệp, nội dung được cá nhân hóa để tăng tương tác, lời khuyên được thiết kế để thúc đẩy hành vi mong muốn. Ở đây, cảm biến không chỉ mở cửa cho máy nhìn thế giới; cảm biến cũng mở cửa cho thế giới - đặc biệt là thế giới của con người - bị nhìn theo cách có thể thương mại hóa và thao túng.

Vì vậy, điều quan trọng nhất cần ghi nhớ là: AI không phải chỉ là một thuật toán nằm trong máy tính, tách rời khỏi đời sống. AI là một chuỗi hoàn chỉnh, một dây chuyền tác động: cảm biến tạo ra dữ liệu; dữ liệu đi qua mô hình; mô hình sinh ra quyết định; quyết định dẫn tới hành động; hành động tạo tác động lên con người và môi trường; tác động lại quay trở lại thành dữ liệu mới. Mỗi mắt xích trong chuỗi này đều có thể mang sai lệch, rủi ro, hoặc bị khai thác. Cảm biến có thể thiên lệch vì đặt sai vị trí, vì chất lượng kém, vì thu thập dữ liệu không đại diện; dữ liệu có thể bị nhiễu, bị làm giả, bị thiếu bối cảnh; mô hình có thể học nhầm tương quan, khuếch đại định kiến, thất bại trong tình huống hiếm; quyết định có thể được tối ưu theo mục tiêu hẹp mà bỏ qua con người; hành động có thể tạo ra hậu quả dây

chuyền; và tác động lên con người có thể trở thành một vòng lặp tự củng cố, nơi hệ thống vừa dự đoán vừa tạo ra thực tại mà nó dự đoán. Khi ta nhìn AI theo chuỗi ấy, ta sẽ bớt bị mê hoặc bởi "một mô hình thông minh", và bắt đầu quan tâm đúng chỗ: thiết kế cảm biến ra sao, kiểm chứng dữ liệu thế nào, đặt mục tiêu tối ưu ở đâu, và xây lớp phòng thủ nào để khi một mắt xích sai, cả dây chuyền không biến sai số nhỏ thành thiệt hại lớn. Chỉ khi hiểu cảm biến như cánh cửa định hình thế giới, ta mới hiểu vì sao "muôn thông minh, phải biết thấy - nghe - chạm - ngửi" - và vì sao trong kỷ nguyên AI, câu hỏi về giác quan của máy cũng đồng thời là câu hỏi về quyền lực của dữ liệu và sự an toàn của con người.

KẾT LUẬN

Nếu bạn đã kiên nhẫn đi hết hành trình của cuốn sách này - từ câu hỏi tưởng như đơn giản nhưng đầy rắc rối "trí tuệ là gì?", đến những hệ AI cổ điển vận hành bằng cấu trúc, ký hiệu và luật lệ; rồi bước sang AI hiện đại học từ dữ liệu, tối ưu bằng xác suất và tính toán; tiếp tục chạm vào thế giới robot, cảm biến, hành động trong môi trường thật - thì đến đây, cảm giác "AI là thứ gì đó bí ẩn, thần bí, khó với tới" sẽ dần tan đi. AI không phải một sinh vật có ý chí, cũng không phải chiếc hộp đen ma thuật luôn đúng. Nó giống một họ công cụ: có cái rất giỏi nhìn ra xu hướng từ hàng triệu dữ kiện, có cái mạnh ở việc lập kế hoạch theo luật, có cái phù hợp để điều khiển máy móc trong thế giới vật lý. Và như mọi công cụ, mỗi loại đều có sức mạnh riêng - đồng thời mang theo những "điểm mù" riêng.

Vì vậy, mục tiêu của việc "hiểu AI" không phải để thuộc lòng thuật ngữ, cũng không phải để chạy theo những từ khóa thời thượng. Thuộc thuật ngữ có thể khiến ta nói chuyện nghe có vẻ chuyên nghiệp, nhưng không giúp ta ra quyết định tốt hơn khi đối diện một hệ thống AI thật đang ảnh hưởng tới công việc, tiền bạc, an toàn, hoặc danh dự của con người. Thứ quan trọng hơn là có một cách nhìn bình tĩnh: biết AI đang làm được gì, không làm được gì; biết lúc nào nên tin, lúc nào phải kiểm tra; biết khi nào nên dùng nó như trợ lý, khi nào phải đặt ranh giới rõ ràng.

AI hiện đại thường tạo cảm giác "thông minh" vì nó trả lời trôi chảy, dự đoán nhanh, gợi ý đúng sở thích, và đôi khi tỏ ra rất "hiểu" con người. Nhưng sự trôi chảy không đồng nghĩa với sự hiểu. Một hệ thống có thể cho ra kết quả đúng nhiều lần chỉ vì nó đã "thấy" quá nhiều mẫu hình tương tự trong dữ liệu, chứ không phải vì nó hiểu bản chất như con người hiểu. Nói cách khác: AI giỏi khớp mẫu hình, giỏi tối ưu mục tiêu đã giao, giỏi xử lý số lượng lớn - nhưng dễ lạc hướng nếu mục tiêu được đặt sai, dữ liệu bị lệch, hoặc môi trường thay đổi theo cách nó chưa từng gặp.

Và đó là lý do cuốn sách này cố gắng dẫn bạn đi qua từng lớp: để khi gặp một ứng dụng AI ngoài đời, bạn không bị choáng bởi lớp vỏ "thần kỳ", cũng không vội phủ định kiểu "AI chỉ là trò quảng cáo". Bạn sẽ nhìn nó như nhìn một cỗ máy: muốn hiểu cỗ máy làm gì thì

phải hỏi nó được thiết kế để làm gì; muốn đánh giá nó đáng tin tới đâu thì phải xem nó dựa vào cái gì; muốn dùng nó an toàn thì phải biết nó thường sai theo kiểu nào.

Từ góc nhìn thực tế, "hiểu AI" rốt cuộc nằm ở việc giữ lại trong đầu vài câu hỏi nền - những câu hỏi đơn giản, dễ nhớ, nhưng đủ sắc để bạn không bị cuốn theo cảm xúc, quảng cáo, hay sự tự tin quá mức của một bản demo. Mỗi lần bạn chuẩn bị dùng AI, mua một giải pháp AI, hoặc cho AI tham gia vào một quy trình quan trọng, hãy tự quay về những câu hỏi này:

Ta đang dùng AI để làm loại việc gì - dự đoán, tối ưu, ra quyết định, hay thay thế một phần tương tác con người? Nếu là dự đoán, ta chấp nhận sai số tới đâu? Nếu là tối ưu, mục tiêu có bị đặt lệch khiến hệ thống "đạt KPI" nhưng gây hại không? Nếu là ra quyết định, ai chịu trách nhiệm khi sai? Nếu là thay thế tương tác con người, ta có đang bỏ quên cảm xúc, đạo đức, sự công bằng, và bối cảnh không?

Hệ thống này mạnh vì dữ liệu hay vì cấu trúc? Có hệ thống mạnh nhờ dữ liệu cực lớn và học thống kê; có hệ thống mạnh nhờ luật rõ ràng và cấu trúc chặt chẽ; cũng có hệ lai. Biết nó "ăn" cái gì sẽ giúp bạn hiểu nó "đói" cái gì: thiếu dữ liệu thì nó bịa theo mẫu; thiếu cấu trúc thì nó lạc trong các trường hợp hiếm; thiếu bối cảnh thì nó trả lời hay nhưng không đúng việc.

Sai lầm thường đến từ đâu - luật sai, dữ liệu lệch, hay môi trường thay đổi? Đây là câu hỏi để chẩn đoán. Nếu luật sai, sửa luật. Nếu dữ liệu lệch, sửa dữ liệu hoặc cách thu thập. Nếu môi trường thay đổi, cần cập nhật mô hình, hoặc thiết kế cơ chế thích nghi. Không có câu hỏi này, ta dễ đổ lỗi mơ hồ kiểu "AI dở", trong khi vấn đề thật là "đầu vào dở" hoặc "bài toán đặt sai".

Khi nó đúng, nó đúng vì "hiểu" hay đúng vì khớp mẫu hình? Câu hỏi này không nhằm chê AI, mà để đặt mức kỳ vọng đúng. Nếu nó chỉ khớp mẫu hình, thì trong tình huống mới lạ, nó có thể sai rất tự tin. Nếu nó có cấu trúc giải thích được (ví dụ dựa trên luật hoặc ràng buộc rõ), thì bạn có thể kiểm soát và dự đoán hành vi của nó tốt hơn. Càng hiểu nguồn gốc của "đúng", bạn càng biết cách dùng nó đúng chỗ.

Khi nó sai, ta có cơ chế phát hiện và giới hạn hậu quả không? Đây là câu hỏi của người làm chủ, không phải người dùng theo cảm tính.

AI sai là chuyện bình thường; điều nguy hiểm là AI sai mà ta không biết nó sai, hoặc biết mà không có rào chắn để giảm thiệt hại. Cơ chế có thể là kiểm tra chéo, cảnh báo độ tự tin, nhật ký, quy trình phê duyệt, quyền hạn giới hạn, hoặc "con người trong vòng lặp" ở những điểm nhạy cảm.

Và khi bạn mang theo những câu hỏi nền này, cuốn sách hoàn thành vai trò của nó: không biến bạn thành nhà nghiên cứu AI, nhưng giúp bạn trở thành người đọc tỉnh táo của thời đại AI - biết nhìn xuyên qua lớp hào quang, biết khen đúng chỗ, biết nghi ngờ đúng lúc, và biết thiết kế cách sử dụng an toàn. Bạn sẽ không còn sợ AI theo kiểu mơ hồ, cũng không còn mê AI theo kiểu mù quáng. Bạn sẽ "tỉnh": tỉnh để hiểu công cụ nào phù hợp với việc gì, tỉnh để không giao nhầm quyền lực cho một hệ thống không xứng đáng, và tỉnh để luôn nhớ rằng công nghệ mạnh đến đâu cũng chỉ là phần phụ - phần chính vẫn là con người: cách ta đặt câu hỏi, đặt mục tiêu, đặt ranh giới, và chịu trách nhiệm cho những gì mình chọn.

“Hiểu AI” không phải để sợ, mà để tỉnh táo