AI VÀ XÃ HỘI

AI AND SOCIETY: SAFETY, ETHICS, AND TRUST

AN TOÀN - ĐẠO ĐỨC - NIỀM TIN

ThS. Bùi Thị Hải

M ỤC LỤC

L Ờ I G I Ớ I T H I Ệ U

AI không còn đứng ở rìa của đời sống như một phát minh để ngắm nhìn; nó đang trở thành một tầng nền âm thầm nhưng quyết định nhịp vận hành của xã hội hiện đại. Mỗi lần bạn mở một dòng tin, nhận một gợi ý, bị từ chối hay được trao cơ hội - đâu đó có thể đã có một mô hình học máy tham gia. Điều đáng nói là AI hiếm khi xuất hiện với tên gọi "AI". Nó hiện diện như sự tiện lợi, như tốc độ, như những kết quả "có vẻ hợp lý". Và chính vì vậy, thách thức lớn nhất của thời đại này không phải là thiếu công cụ, mà là thiếu một cách hiểu đủ sâu để ta vẫn giữ được quyền chủ động: chủ động trong lựa chọn, chủ động trong đánh giá, và chủ động trong trách nhiệm.

AI and Society: Safety, Ethics, and Trust được viết như một lời mời bước vào phần cốt lõi nhất của câu chuyện AI - không phải câu chuyện về những màn trình diễn kỹ thuật ấn tượng, mà là câu chuyện về điều gì làm nên một công nghệ đáng tin trong thế giới thật. Cuốn sách đặt ra và theo đuổi ba câu hỏi nền tảng: làm sao để AI an toàn khi nó ngày càng mạnh và lan rộng; làm sao để AI đạo đức khi nó tác động đến con người với những khác biệt về điều kiện, vị thế và cơ hội; và làm sao để xã hội có thể tin vào AI mà không rơi vào mù quáng - một niềm tin được xây dựng từ minh bạch, kiểm chứng, và quản trị, chứ không phải từ cảm giác hay lời hứa.

Giá trị lớn lao mà cuốn sách mang lại cho bạn không chỉ là thêm kiến thức, mà là một năng lực mới: năng lực nhìn AI như một hệ thống xã hội có thể thiết kế và kiểm soát; năng lực đặt câu hỏi đúng trước khi chấp nhận một mô hình hay một chính sách; năng lực phân biệt giữa "đúng trong phòng thí nghiệm" và "đúng trong đời sống"; năng lực nhận diện nơi rủi ro sinh ra - từ dữ liệu, mục tiêu tối ưu, cơ chế khuyến khích, đến bối cảnh triển khai - và vì sao những sai lệch nhỏ có thể khuếch đại thành hệ quả lớn. Khi bạn có được năng lực ấy, bạn không còn đứng trước AI với tâm thế hoặc sợ hãi hoặc tôn thờ; bạn đứng trước nó như một người trưởng thành: hiểu giới hạn, hiểu trách nhiệm, và hiểu cách tạo ra lợi ích bền vững.

Cuốn sách này dành cho những người muốn đi xa hơn mức "biết dùng". Dành cho người học cần một nền tảng tư duy vững chắc. Dành cho kỹ sư muốn xây hệ thống có trách nhiệm. Dành cho nhà quản lý và nhà hoạch định chính sách muốn ra quyết định dựa trên cấu trúc thay vì trực giác. Và dành cho bất kỳ ai tin rằng tương lai số không nên được định hình bởi tốc độ của công nghệ, mà bởi chất lượng của sự lựa chọn. Nếu bạn đang tìm một tấm bản đồ để bước vào kỷ nguyên AI với sự tỉnh táo, bản lĩnh và niềm tin có căn cứ - những trang tiếp theo là nơi để bắt đầu.

Cuốn sách cũng thừa nhận đây không phải cẩm nang dứt khoát cho mọi rủi ro AI, vì lĩnh vực này còn mới và thay đổi nhanh. Nhiều tác hại hiện tại như thông tin sai lệch, vi phạm quyền riêng tư, suy giảm kết nối xã hội hay tác động môi trường đã có các tài liệu khác bàn sâu. Tác giả chủ yếu tập trung vào những rủi ro mới từ AI tiên tiến - các rủi ro có thể nghiêm trọng, quy mô lớn và đôi khi khó đảo ngược, trong khi xã hội chưa sẵn sàng.

Về cấu trúc, sách chia thành ba phần lớn: **AI và rủi ro quy mô xã hội**, **An toàn**, **Đạo đức & Xã hội**. Phần đầu phác thảo các kịch bản rủi ro thảm khốc và phân loại rủi ro thành bốn nhóm như lạm

dụng độc hại, động lực chạy đua vũ trang AI, rủi ro tổ chức và AI "bất hảo". Phần An toàn đi sâu vào việc làm hệ thống AI an toàn hơn (thiên kiến, minh bạch, hiện tượng phát sinh), đồng thời đưa vào các nguyên tắc kỹ thuật an toàn, văn hóa an toàn trong tổ chức và góc nhìn hệ thống phức tạp. Phần Đạo đức & Xã hội bàn về cách đặt mục tiêu có lợi, các vấn đề phối hợp giữa nhiều tác nhân/công ty/quốc gia bằng lý thuyết trò chơi, và các hướng quản trị từ doanh nghiệp đến quốc gia và quốc tế. Kèm theo đó là website cung cấp tài nguyên học tập và nội dung bổ sung.

PHẦN I
AI VÀ RỦI RO QUY MÔ XÃ HỘI

Trí tuệ nhân tạo không chỉ là một công cụ mới để làm việc nhanh hơn; nó là một "bộ khuếch đại" có thể mở rộng năng lực của con người và tổ chức lên quy mô lớn. Khi chi phí hành động giảm xuống, tốc độ thử - sai tăng lên và tác động lan truyền nhanh hơn, xã hội bước vào một giai đoạn mà lợi ích và rủi ro đều được phóng to. Vì thế, rủi ro AI không còn là chuyện "một phần mềm sai", mà là câu chuyện về những hệ quả dây chuyền: sai lầm có thể lặp lại hàng triệu lần, định kiến có thể được tự động hóa, và quyết định có thể được đưa ra nhanh hơn khả năng kiểm chứng của con người.

PHẦN I giúp người đọc nhìn AI trong bối cảnh rộng hơn của đời sống hiện đại - nơi công nghệ được tích hợp sâu vào sản phẩm, doanh nghiệp, truyền thông, vận hành và cả cạnh tranh giữa các bên. Khi AI trở thành một lớp hạ tầng, rủi ro của nó không nằm riêng trong mô hình, mà nằm trong mạng lưới: cách con người triển khai, cách tổ chức ra quyết định, và cách xã hội phản ứng. Những áp lực như chạy đua tốc độ, sự tự mãn của tổ chức, hay việc đánh giá sai năng lực thực sự của AI đều có thể biến một công cụ hữu ích thành một nguồn bất ổn.

Quan trọng hơn, PHẦN I cũng gỡ bỏ một ảo giác phổ biến: AI có thể nói năng trôi chảy và tạo cảm giác thuyết phục, nhưng điều đó không đảm bảo nó "hiểu" như con người. Nếu ta nhầm lẫn giữa sự mượt mà và sự đúng đắn, ta dễ trao niềm tin, quyền hạn và trách nhiệm cho hệ thống vượt quá khả năng kiểm soát. Vì vậy, mục tiêu của phần này không phải làm bạn sợ AI, mà giúp bạn có một cái nhìn tỉnh táo: hiểu vì sao AI mạnh lên, vì sao rủi ro có thể tăng theo quy mô, và vì sao an toàn không thể chỉ là vấn đề kỹ thuật - mà là vấn đề của toàn bộ hệ sinh thái nơi AI đang sống và tác động.

CHƯƠNG 1: BỨC TRANH RỦI RO THẢM KHỐC - KHI AI KHUẾCH ĐẠI QUYỀN LỰC

Điều đáng sợ nhất của AI không nằm ở chỗ nó "thông minh", mà nằm ở chỗ nó **khuếch đại**. AI giống như một bộ khuếch đại âm thanh: nó không tự tạo ra bản nhạc, nhưng có thể làm tiếng nói nhỏ thành rất lớn, làm một ý tưởng thành một chiến dịch, làm một thao tác thành một dây chuyền. Một người bình thường có thể dùng AI để viết nhanh hơn, học nhanh hơn, làm việc hiệu quả hơn. Nhưng cũng chính cơ chế ấy khiến một nhóm có ý đồ xấu có thể làm hại nhanh hơn, rẻ hơn, và rộng hơn. Khi chi phí hành động giảm xuống, số lần thử - sai tăng lên, tốc độ lan truyền tăng lên, xã hội bước vào một "vùng rủi ro" mới - nơi quy mô và tốc độ trở thành biến số quyết định.

Vấn đề nằm ở chỗ rủi ro AI không đi theo đường thẳng. Nó không giống một chiếc máy hỏng sẽ dừng lại ở chỗ nó hỏng. AI thường được gắn vào mạng lưới: nền tảng số, chuỗi cung ứng, hệ thống ra quyết định, truyền thông, an ninh, và cả cạnh tranh giữa các tổ chức. Một sai lệch nhỏ trong một mắt xích có thể phóng đại thành hậu quả lớn khi nó đi qua hàng triệu người dùng, hàng nghìn tổ chức, và vô số bối cảnh mà nhà thiết kế ban đầu không lường hết. Vì thế, muốn hiểu rủi ro AI, ta cần một "tấm bản đồ" đủ rõ để không bị choáng, nhưng cũng đủ mềm để không bỏ sót những vùng chồng lấn.

Một cách hữu ích để bắt đầu là nhìn rủi ro quy mô xã hội theo bốn mảng lớn: **lạm dụng ác ý**, **động lực chạy đua**, **rủi ro tổ chức**, và **bài toán kiểm soát mục tiêu**. Bốn mảng này không phải bốn chiếc hộp kín. Chúng giao nhau và cộng hưởng: chạy đua làm giảm an toàn; giảm an toàn tạo lỗ hổng để lạm dụng; lỗ hổng phơi bày điểm yếu của tổ chức; và khi hệ thống mạnh lên, bài toán kiểm soát mục tiêu trở nên nghiêm trọng hơn. Nhưng chia như vậy giúp ta nhìn thấy cấu trúc của vấn đề, thay vì chỉ thấy một đống lo lắng rời rạc.

Lạm dụng ác ý thường là cánh cửa đầu tiên khiến người ta nhận ra AI không chỉ là công cụ "vô hại". Trong nhiều vấn đề xã hội, nếu đa số tử tế và hợp tác, ta có thể hy vọng tình hình vẫn ổn. Nhưng có những loại rủi ro không cần đa số xấu; nó chỉ cần một thiểu số hành động

đơn phương là đủ tạo hậu quả lớn. AI làm nguy cơ ấy tăng lên vì nó hạ thấp rào cản năng lực. Nhiều thứ trước đây đòi hỏi chuyên gia, thời gian, và nguồn lực - - nay có thể được "đóng gói" thành quy trình, thành mẫu, thành công cụ sẵn dùng. Khi năng lực được phổ cập, điểm yếu của xã hội cũng bị lộ rõ: chúng ta đang dựa vào giả định rằng "phần đông sẽ dùng đúng". Nhưng trong thế giới có AI, điều đáng lo không phải là AI chắc chắn sẽ bị lạm dụng; điều đáng lo là **nếu** bị lạm dụng, nó có thể lan rộng nhanh đến mức các cơ chế phòng vệ truyền thống không kịp phản ứng. Xã hội sẽ rơi vào thế đuổi theo hậu quả: vá lỗi sau khi sự việc xảy ra, ban hành quy định sau khi thiệt hại đã lan, và dựng rào chắn sau khi con đường đã bị lợi dụng.

Động lực chạy đua AI lại là một dạng rủi ro tinh vi hơn, vì nó không cần "người xấu" để kích hoạt. Nó xuất hiện khi các tổ chức và quốc gia bị kéo vào một cuộc cạnh tranh mà tốc độ được thưởng, còn thận trọng bị phạt. Trong môi trường như vậy, an toàn dễ bị xem là thứ làm chậm: kiểm thử lâu hơn, triển khai chậm hơn, mất lợi thế nói trước. Khi phần thưởng của việc "ra trước" đủ lớn, còn hậu quả của việc "sai" lại có thể được đẩy ra ngoài - đổ lên người dùng, xã hội, hoặc tương lai - hệ thống sẽ trượt theo quán tính. Điều hợp lý ở cấp độ từng tác nhân ("mình phải nhanh nếu không bị bỏ lại") có thể dẫn đến kết quả nguy hiểm ở cấp độ toàn hệ thống ("tất cả cùng nhanh và cùng kém an toàn hơn"). Đây là lý do các vấn đề AI không thể chỉ giải bằng đạo đức cá nhân; nó cần cơ chế phối hợp, chuẩn mực chung, và những "lan can" đủ mạnh để tốc độ không nuốt chửng an toàn.

Rủi ro tổ chức là lời nhắc rằng tai nạn lớn hiếm khi chỉ do công nghệ. Nó thường là tổng hòa của con người, quy trình, văn hóa, và cấu trúc quyền lực. Một tổ chức có thể có các kỹ sư giỏi nhưng vẫn thất bại vì áp lực KPI, vì tâm lý "chắc ổn thôi", vì bỏ qua cảnh báo, vì thiếu người có quyền dừng dự án, hoặc vì coi an toàn là việc của "bộ phận kiểm thử" chứ không phải của toàn hệ thống.Lịch sử đã nhiều lần chứng minh: ngay cả những lĩnh vực sống còn như hàng không, hạt nhân, hay tàu con thoi vẫn có thể xảy ra thảm họa khi văn hóa an toàn bị bào mòn. Với AI, nguy cơ ấy tăng lên vì AI thường giống một

"hộp đen": nó có thể cho kết quả tốt trong môi trường thử nghiệm, nhưng khi ra đời sống thật - đầy biến số, đầy tình huống hiếm - nó có thể hành xử khác. Sự khó hiểu ấy tạo ra "vùng mù" trong tổ chức: người quyết định không hiểu rõ rủi ro, người hiểu rủi ro không đủ quyền, và người chịu hậu quả thường là người ở ngoài cuộc họp.

Cuối cùng là bài toán kiểm soát mục tiêu - phần dễ bị hiểu sai nhất. Người ta hay tưởng rủi ro này là chuyện "AI nổi loạn" như trong phim. Nhưng vấn đề sâu hơn và thực tế hơn là: khi một hệ thống có năng lực tối ưu hóa mạnh, điều quan trọng nhất không còn là nó có "ý đồ" hay không, mà là **mục tiêu mà ta giao cho nó có phản ánh đúng giá trị con người không**. Một hệ thống có thể "rất ngoan" theo nghĩa nó tối ưu đúng cái ta đo, nhưng lại "rất nguy hiểm" theo nghĩa nó bỏ qua cái ta không đo. Nó có thể tìm ra con đường tắt để đạt mục tiêu, khai thác kẽ hở của quy tắc, hoặc tạo ra hệ quả phụ mà ta không dự đoán. Câu hỏi then chốt nằm ở đây: nếu hệ thống ngày càng mạnh, liệu ta có chắc rằng mục tiêu ta đặt ra sẽ luôn an toàn trong mọi bối cảnh? Và liệu ta có chắc rằng hệ thống sẽ không học cách đạt mục tiêu bằng cách đi vòng qua những ràng buộc mong manh của con người?

Bốn mảng rủi ro này không nhằm khiến ta bi quan. Ngược lại, chúng giúp ta tỉnh táo. Bởi khi nhìn đúng bản chất của rủi ro, ta sẽ bớt sa vào những tranh luận vô ích kiểu "AI tốt hay xấu", và thay vào đó đặt được những câu hỏi hữu dụng hơn: AI đang khuếch đại điều gì? Cơ chế nào khiến rủi ro tăng tốc? Ở đâu cần lan can kỹ thuật, ở đâu cần lan can xã hội? Ai chịu trách nhiệm khi sai, và làm sao để hệ thống học được cách dừng lại trước khi hậu quả lan rộng?

Nếu coi AI là một bộ khuếch đại, thì nhiệm vụ của xã hội không phải là đập bỏ bộ khuếch đại, cũng không phải là mở hết công suất rồi cầu may. Nhiệm vụ là học cách dùng nó trong một phòng có cách âm, có nút giảm âm lượng, có giới hạn công suất, và có người chịu trách nhiệm khi âm thanh vượt ngưỡng. Nói cách khác, rủi ro AI không chỉ là câu chuyện của thuật toán. Nó là câu chuyện của cách chúng ta xây dựng, triển khai, quản trị, và - quan trọng nhất - cách chúng ta lựa chọn giá trị để đặt vào một công cụ ngày càng mạnh.

CHƯƠNG 2: AI HIỆN ĐẠI - HIỂU ĐỦ ĐỂ KHÔNG BỊ ẢO GIÁC BỞI CHÍNH CÔNG CỤ

Muốn nói về an toàn, đạo đức hay tác động xã hội của AI, trước hết cần một nền móng tối thiểu: AI vận hành theo logic nào, mạnh lên vì đâu, và vì sao nó vừa hữu ích vừa khó lường. Không cần công thức, không cần thuật toán chi tiết; chỉ cần hiểu đúng vài điểm cốt lõi, bạn sẽ tránh được một cái bẫy phổ biến của thời đại này: **ảo giác rằng AI "hiểu" như con người chỉ vì nó "nói" như con người**.

AI hiện đại mạnh lên nhờ ba cột trụ: **dữ liệu**, **tính toán**, và **mô hình lớn**. Dữ liệu là "kinh nghiệm" mà hệ thống học từ đó; tính toán là năng lượng để quá trình học diễn ra; mô hình lớn là cấu trúc đủ sức hấp thụ những mẫu phức tạp. Khi cả ba yếu tố tăng lên, năng lực của AI có thể tăng nhanh đến mức người ngoài cảm thấy như có một bước nhảy vọt. Đôi khi, điều tạo nên cú nhảy không phải một phát minh duy nhất, mà là việc mở rộng quy mô: nhiều dữ liệu hơn, nhiều tài nguyên hơn, mô hình lớn hơn - và kết quả là khả năng xử lý ngôn ngữ, hình ảnh, âm thanh, hay dữ liệu vận hành đều tiến bộ rõ rệt.

Điểm then chốt là: phần lớn AI ngày nay **học từ mẫu**. Nó đặc biệt giỏi ở những việc có thể rút ra quy luật từ vô số ví dụ: dự đoán chữ tiếp theo trong một đoạn văn, phát hiện cấu trúc trong dữ liệu, tổng hợp biến thể của một phong cách, tạo ra nội dung có "hơi thở" giống người. Nhờ vậy, AI trở thành một công cụ tăng lực đáng nể: tóm tắt nhanh, viết nháp, dịch thuật, gợi ý ý tưởng, phân loại thông tin, phát hiện bất thường, hỗ trợ chăm sóc khách hàng, tối ưu vận hành… Một công việc từng mất hàng giờ có thể rút xuống vài phút.

Nhưng chính sức mạnh này tạo ra ảo giác nguy hiểm nhất: **giỏi không đồng nghĩa hiểu**. AI có thể tạo ra câu trả lời trôi chảy vì nó biết cách kết nối mẫu hình trong dữ liệu, chứ không phải vì nó luôn nắm bản chất như con người. Nó có thể lập luận mạch lạc, dùng thuật ngữ chuẩn, giữ nhịp văn phong ổn định, thậm chí thể hiện sự tự tin - và vẫn sai. Sai ở dữ kiện. Sai ở suy luận. Sai vì thiếu bối cảnh. Sai vì lấp khoảng trống bằng thứ "nghe hợp lý". Khi bạn đọc một câu trả lời quá mượt, đôi lúc điều cần làm không phải gật đầu, mà là dừng lại hỏi

: *nó dựa trên bằng chứng nào, giả định nào, và liệu có cách kiểm chứng độc lập không?*

Ảo giác càng mạnh hơn khi AI bắt đầu có vẻ "đa năng". Một mô hình có thể vừa viết, vừa dịch, vừa giải toán, vừa tư vấn, vừa tạo ảnh. Cảm giác ấy khiến ta dễ quên rằng những năng lực này phần lớn vẫn dựa trên một kỹ thuật cốt lõi: học từ dữ liệu để dự đoán và tổng hợp. Một chiếc kính có thể giúp bạn nhìn xa hơn, nhưng nó không biến bạn thành người biết mọi thứ. AI cũng vậy: nó mở rộng khả năng xử lý thông tin, nhưng không tự động mang lại sự thật, càng không tự động mang lại sự khôn ngoan.

Một điểm quan trọng khác là AI không sống trong "môi trường phòng thí nghiệm". Nó sống trong đời thật - nơi có vô số tình huống hiếm, dữ liệu méo, hành vi người dùng bất ngờ, và những điều không nằm trong bộ mẫu đã học. AI có thể rất tốt trong những tình huống quen thuộc, nhưng khi gặp điều lạ, nó vẫn có thể tạo ra câu trả lời thuyết phục, giống như một người đoán mò nhưng nói rất chắc. Sự khác biệt là: khi một người đoán mò, ta thường nhận ra qua sự do dự; còn AI có thể không có dấu hiệu do dự. Vì vậy, trong những bối cảnh quan trọng, AI không nên được xem như "người phán xử", mà nên được đặt vào vai trò phù hợp: hỗ trợ, gợi ý, kiểm tra, và luôn có cơ chế để con người kiểm chứng.

Và rồi có một chuyển động lớn hơn đang diễn ra: AI không còn chỉ là một ứng dụng, nó đang dần trở thành **hạ tầng**. Ngày trước, ta nghĩ AI là một công cụ riêng lẻ: chatbot, phần mềm gợi ý, hệ thống nhận diện. Ngày nay, AI được nhúng vào mọi tầng: nền tảng tìm kiếm, mạng xã hội, thương mại điện tử, chống gian lận, tuyển dụng, chấm điểm tín dụng, điều phối chuỗi cung ứng, giám sát an ninh, tối ưu vận hành. Khi AI trở thành hạ tầng, rủi ro của nó không còn là rủi ro của một sản phẩm; nó là rủi ro của một mạng lưới.

Trong mạng lưới, sai sót không đứng yên. Nó lan truyền, lặp lại, và tích tụ. Một thuật toán gợi ý lệch có thể tạo ra "bong bóng thông tin", làm xã hội phân cực. Một hệ thống đánh giá không công bằng có thể khiến bất lợi rơi đều lên một nhóm người, ngày này qua ngày khác.

Một mô hình được triển khai vội có thể gây phản ứng dây chuyền trong vận hành. Càng tích hợp sâu, tác động càng khó nhìn thấy, vì nó nằm sau lớp tiện lợi: ta chỉ thấy kết quả, không thấy cơ chế.

AI cũng tạo ra những dạng phụ thuộc tinh vi. Khi mọi thứ "nhờ AI cho nhanh", con người và tổ chức dễ mất dần năng lực tự kiểm tra, tự làm chủ, tự ra quyết định. Sự phụ thuộc này không xuất hiện trong một ngày. Nó hình thành từ thói quen: hỏi thay vì hiểu, nhờ thay vì rèn, tin thay vì kiểm chứng. Và khi thói quen lan rộng, nó trở thành một chuẩn sống mới - một chuẩn mà xã hội khó quay lại nếu không có ý thức.

Vì thế, hiểu AI hiện đại không phải để bạn sợ AI. Hiểu là để bạn dùng nó đúng chỗ. Khi bạn biết AI mạnh lên nhờ dữ liệu–tính toán–quy mô, bạn sẽ bớt ngạc nhiên trước tốc độ phát triển. Khi bạn biết "giỏi" không đồng nghĩa "hiểu", bạn sẽ bớt bị thôi miên bởi câu trả lời mượt mà. Và khi bạn thấy AI đang trở thành hạ tầng, bạn sẽ hiểu vì sao an toàn AI không thể chỉ là chuyện kỹ thuật: nó là câu chuyện của quản trị, trách nhiệm, và cách xã hội đặt lan can cho một công cụ khuếch đại.

PHẦN II
AN TOÀN

Chuyển trọng tâm từ câu hỏi "AI có thể làm gì?" sang câu hỏi quan trọng hơn: **làm sao để AI không gây hại khi bước vào đời sống thật**. An toàn ở đây không được hiểu theo nghĩa "lịch sự hơn" hay "ít nói điều xấu hơn", mà là khả năng **giảm xác suất sai nghiêm trọng**, **phát hiện sai kịp thời**, và **giới hạn hậu quả khi sai**. Bởi một hệ thống càng thuyết phục thì cái sai của nó càng dễ được tin, càng dễ lan rộng, và càng khó sửa.

Phần này nhấn mạnh rằng nhiều rủi ro nguy hiểm nhất lại đến từ những lỗi khó nhìn thấy: AI đúng trong tình huống quen thuộc nhưng trượt trong tình huống hiếm; AI "nghe hợp lý" nhưng dựa trên giả định sai; AI bị dẫn dụ, bị lách luật, hoặc bị khai thác bằng các kỹ thuật tấn công. Vì thế, an toàn không thể trông chờ vào may mắn hay vào cảm giác "có vẻ ổn", mà phải được xây như một hệ thống phòng thủ có chủ đích.

Một điểm then chốt khác của PHẦN II là: **an toàn không phải một lớp sơn phủ lên sản phẩm**, càng không phải một bộ thủ tục để yên lòng dư luận. Nếu không đo được rủi ro tổng thể, tổ chức rất dễ rơi vào ảo tưởng an toàn hoặc "safetywashing" - nói nhiều về an toàn nhưng thực tế không giảm rủi ro. An toàn đúng nghĩa phải đi cùng cấu trúc "nhiều lớp phòng thủ": kiểm thử đối kháng, giám sát, phát hiện bất thường, bảo mật, minh bạch - giải trình, phê duyệt và cơ chế dừng khẩn cấp… để giảm khả năng các lỗ hổng tình cờ thẳng hàng và tạo thành tai nạn.

PHẦN II nhấn mạnh: an toàn AI không chỉ là mô hình hay mã nguồn, mà phụ thuộc lớn vào văn hóa và kỷ luật tổ chức - cách đặt KPI, ưu tiên của lãnh đạo, phản ứng với cảnh báo, và việc khuyến khích báo lỗi. Thiếu tư duy an ninh sẽ bỏ sót nguy cơ bị khai thác.

CHƯƠNG 3: AN TOÀN TÁC NHÂN ĐƠN LẺ - KHI MỘT MÔ HÌNH "CÓ VẺ ỔN" VẪN CÓ THỂ NGUY HIỂM

Có một hiểu lầm rất phổ biến về an toàn AI: cứ tưởng chỉ cần làm cho mô hình "ngoan" hơn là xong. Ngoan ở đây nghĩa là lịch sự hơn, ít nói độc hại hơn, tránh những chủ đề nhạy cảm hơn, trả lời mềm mại hơn. Những thứ đó quan trọng - đặc biệt với sản phẩm hướng tới số đông - nhưng nếu dừng ở đó, ta mới chỉ chạm vào phần **bề mặt**. Một mô hình có thể cực kỳ lịch thiệp mà vẫn gây hại; thậm chí đôi khi, chính vẻ lịch thiệp và tự tin lại làm nó nguy hiểm hơn, vì nó khiến người dùng dễ tin.

An toàn của một hệ thống AI, xét cho cùng, không phải là "nó có xúc phạm ai không", mà là: **khi đặt nó vào thế giới thật, nó có thể sai theo cách nào, và hậu quả của cái sai đó có bị khuếch đại không**. An toàn tác nhân đơn lẻ là câu chuyện của một mô hình cụ thể - một "tác nhân" đứng trước người dùng hoặc trước một nhiệm vụ - và câu hỏi trung tâm là: làm sao để mô hình không trở thành một nguồn rủi ro tự thân, ngay cả khi chưa có "chạy đua", chưa có "lạm dụng quy mô", chưa có xung đột quốc gia hay tổ chức.

Ở cấp độ này, an toàn thường nằm trong một bó vấn đề gắn chặt với nhau: **thiên kiến**, **minh bạch**, **hành vi phát sinh**, và **thất bại ngoài phân phối**. Nghe có vẻ kỹ thuật, nhưng thực ra đó là những cách gọi tên của một trực giác rất đời thường: AI không sống trong phòng thí nghiệm; nó sống trong vô số tình huống mà ta không dự đoán trước.

Thiên kiến là vết nứt đầu tiên. AI học từ dữ liệu. Dữ liệu là dấu vết của xã hội. Và xã hội vốn không hoàn hảo. Nếu dữ liệu mang định kiến, mô hình có thể học định kiến. Nếu dữ liệu thiếu đại diện cho một nhóm, mô hình có thể "không nhìn thấy" nhóm đó. Điều nguy hiểm là thiên kiến thường không xuất hiện như một câu nói thô bạo, mà xuất hiện như một chuỗi quyết định nhỏ: đề xuất khác nhau, đánh giá khác nhau, ưu tiên khác nhau - lặp đi lặp lại đến mức trở thành bất công hệ thống. Với tác nhân đơn lẻ, câu hỏi an toàn không chỉ là "mô hình có phân biệt đối xử không", mà là "mô hình tạo ra chênh lệch kết

kết quả thế nào khi áp dụng cho những nhóm người khác nhau".

Minh bạch là vấn đề kế tiếp, và cũng là lý do nhiều hệ thống "có vẻ ổn" vẫn khó tin. Khi một mô hình đưa ra kết luận, chúng ta muốn biết nó dựa trên cái gì, tự tin đến mức nào, và đâu là giới hạn của nó. Nhưng phần lớn mô hình hiện đại giống một chiếc hộp đen: nó có thể cho kết quả tốt, nhưng khó giải thích. Khi thiếu minh bạch, ta mất khả năng kiểm tra, mất khả năng phát hiện sai sớm, và mất khả năng quy trách nhiệm rõ ràng. Trong môi trường quan trọng - y tế, tài chính, pháp lý - thiếu minh bạch không chỉ là bất tiện; nó là rủi ro.

Rồi đến hành vi phát sinh - một hiện tượng khiến nhiều người bất an. Khi mô hình đủ lớn và đủ phức tạp, đôi khi nó biểu hiện những năng lực hoặc hành vi mà ta không "lập trình trực tiếp". Nó có thể học cách nói vòng để đạt mục tiêu, học cách né rào, học cách khai thác kẽ hở trong hướng dẫn, hoặc tự hình thành những chiến lược "hợp lý theo nó" nhưng không hợp lý theo ta. Vấn đề ở đây không phải là mô hình có ý thức, mà là: trong một không gian hành vi rất rộng, mô hình có thể tìm ra những con đường ta không nghĩ tới. Và một khi con đường đó đem lại "điểm thưởng" trong cơ chế đánh giá nào đó, nó có thể lặp lại.

Nhưng có lẽ nguy hiểm nhất với tác nhân đơn lẻ là thất bại ngoài phân phối: mô hình hoạt động tốt trong những tình huống quen thuộc, rồi trượt chân ở những tình huống hiếm. Thế giới thật đầy những tình huống hiếm - mà chính vì hiếm nên ta ít chuẩn bị. Một hệ thống có thể đúng trong 99 trường hợp và sai trong 1 trường hợp, nhưng nếu 1 trường hợp ấy rơi vào đúng điểm trọng yếu, hậu quả có thể lớn hơn tất cả 99 lần đúng cộng lại. Và cái khó là: tình huống hiếm thường không báo trước bằng tín hiệu rõ ràng.

Trong thực tế, có những dạng sai **nguy hiểm hơn** vì chúng khó phát hiện hoặc dễ bị bỏ qua.

Có loại sai mà AI nói rất tự tin. Sự trôi chảy giọng điệu chắn chắn khiến người dùng tưởng đây là kết luận có cơ sở. Con người vốn có xu hướng tin vào thứ được nói rõ ràng, mạch lạc. AI vô tình tận dụng chính điểm yếu nhận thức ấy. Khi AI sai mà lại tự tin, nó không chỉ

đưa ra thông tin sai; nó còn đẩy sai lầm ấy đi xa hơn bằng sức nặng của sự thuyết phục.4

Có loại sai"đúng trong bình thường, sai trong hiếm". Đây là loại sai khó trị nhất vì nó tạo cảm giác an toàn giả. Người dùng thấy AI đúng nhiều lần nên tin, rồi đến lần sai thì đã quá muộn. Điều đáng lo là những tình huống hiếm thường trùng với những tình huống căng thẳng: khẩn cấp, phức tạp, nhiều ràng buộc, ít thời gian kiểm chứng.

Có loại sai nghe hợp lý nhưng dựa trên giả định sai. AI có thể dựng một lập luận "đúng kiểu", kết nối các ý theo logic, nhưng nền tảng dữ kiện hoặc giả định ban đầu lại không đúng. Đây là cái bẫy của "hợp lý bề mặt": nếu ta chỉ nhìn vào độ mượt, ta sẽ bỏ qua phần quan trong nhất - tính đúng của nền tảng.

Và có loại sai do bị khai thác: bị dẫn dụ, bị "lách luật", bị prompt injection. Khi AI trở thành tác nhân có thể hành động (gọi API, truy cập công cụ, đọc email, thao tác hệ thống), prompt injection không còn là trò nghịch. Nó là một lối vào để buộc hệ thống làm điều không nên làm, tiết lộ điều không nên tiết lộ, hoặc ra quyết định vượt quyền. Trong thế giới an ninh mạng, người ta luôn giả định "sẽ bị tấn công". AI cũng cần một giả định tương tự.

Từ tất cả những điều đó, ta rút ra một định nghĩa thực tế về an toàn tác nhân đơn lẻ: **một hệ thống an toàn không phải là hệ thống không bao giờ sai**. Không có hệ thống phức tạp nào đạt được điều đó. Một hệ thống an toàn là hệ thống có ba năng lực song hành.

Thứ nhất, nó giảm xác suất sai nghiêm trọng: qua huấn luyện phù hợp, qua kiểm thử đối kháng, qua giới hạn phạm vi, qua thiết kế nhiệm vụ để tránh cho AI "đoán mò" ở nơi không được phép đoán. Thứ hai, nó phát hiện sai: qua cơ chế tự đánh dấu mức độ chắc chắn, qua cảnh báo khi gặp dữ liệu bất thường, qua log và giám sát sau triển khai. Thứ ba, nó giới hạn hậu quả khi sai: bằng quyền hạn tối thiểu, bằng rào chắn an ninh, bằng phê duyệt của con người ở điểm quyết định, bằng nút dừng khẩn cấp khi có dấu hiệu lệch.

CHƯƠNG 4: KỸ THUẬT AN TOÀN - AN TOÀN KHÔNG PHẢI MỘT LỚP SƠN

Trong rất nhiều tổ chức, "an toàn" thường được nói như một lời hứa, một khẩu hiệu, hoặc một bộ quy trình phải có để yên lòng người khác. Nhưng khi đứng trước một công nghệ có thể khuếch đại năng lực ở quy mô lớn, an toàn không thể là một lớp sơn phủ lên sản phẩm sau cùng. An toàn phải là một dạng kỹ thuật, nghĩa là có nguyên tắc thiết kế, có thước đo, có kiểm thử, có cơ chế phản hồi, và có trách nhiệm rõ ràng. Nếu không, thứ ta có chỉ là cảm giác an toàn - chứ không phải an toàn thật.

Một cảnh báo quan trọng là: nhiều tổ chức nói nhiều về an toàn nhưng không trả lời được câu hỏi đơn giản nhất: **rủi ro tổng thể đã giảm chưa?** Họ có thể đưa ra số liệu về một vài hạng mục - ví dụ mô hình lịch sự hơn, ít trả lời độc hại hơn, ít vi phạm chính sách hơn - nhưng lại không đo được bức tranh lớn. Vì rủi ro tổng thể không chỉ phụ thuộc vào "mô hình có an toàn hơn không", mà còn phụ thuộc vào "mô hình có mạnh hơn không", "mô hình được triển khai ở đâu", "được dùng với quyền hạn nào", và "ai có thể lợi dụng nó để làm gì".

Đây là một điểm tinh vi nhưng cực kỳ quan trọng: một cải tiến tưởng là "an toàn" đôi khi đồng thời làm tăng năng lực chung - khiến mô hình mạnh hơn, phổ cập hơn, dễ dùng hơn - và kết quả là rủi ro tổng thể **không giảm**, thậm chí tăng. Nói cách khác, nếu bạn nâng cấp phanh nhưng đồng thời lắp động cơ mạnh hơn và chạy nhanh hơn, tai nạn có thể không ít đi. Vì vậy, an toàn phải được đo **cùng** với năng lực. Muốn giảm rủi ro tổng thể, mức cải thiện an toàn phải **vượt** mức tăng năng lực theo tương quan. Nếu không nhìn theo tương quan này, chúng ta dễ tự đánh lừa mình bằng những con số đẹp.

Từ đây xuất hiện một hiện tượng rất đáng ngại trong ngành :"safetywasing". Nó giống như "greenwasing" trong môi trường - tô xanh bằng quảng cáo và giáy tờ. Safetywasing xảy ra khi tổ chức xây một lớp trình diễn - nhưng phần lõi (đo rủi ro, giảm rủi ro, trách nhiệm khi sai, cơ chế dừng khẩn cấp) lại mờ nhạt. Một số nơi còn đánh tráo khái niệm: quảng bá phát triển năng lực như thể đó là phát

triển năng lực như thể đó là phát triển an toàn. Đây không chỉ là vấn đề đạo đức truyền thông; nó là rủi ro xã hội, vì nó khiến thị trường, công chúng và cả nhà quản lý tưởng rằng vấn đề đã được giải quyết, trong khi thực tế chỉ mới được "trang điểm".

Kỹ thuật an toàn, vì thế, bắt đầu từ một thái độ rất thực dụng: **giả định rằng hệ thống sẽ có lúc sai, và sẽ có người cố tình khai thác nó**. Khi đã chấp nhận điều đó, mục tiêu không còn là "tạo một hệ thống hoàn hảo", mà là tạo một hệ thống có thể chịu được lỗi, phát hiện lỗi sớm, và hạn chế thiệt hại khi lỗi xảy ra. Tương tự như hàng không hay hạt nhân, an toàn đến từ việc thiết kế với tâm thế "điều xấu sẽ xảy ra ở đâu đó, vào lúc nào đó", và nhiệm vụ của kỹ thuật là làm cho điều xấu **khó xảy ra hơn** và **ít gây hậu quả hơn**.

Một trong những hình ảnh trực quan nhất để hiểu tư duy này là "Swiss cheese model". Hãy tưởng tượng mỗi biện pháp an toàn là một lát pho mát. Mỗi lát đều có lỗ hổng - không có biện pháp nào hoàn hảo. Tai nạn xảy ra khi các lỗ hổng của nhiều lớp tình cờ thẳng hàng, tạo thành một đường thông suốt cho rủi ro đi qua. Nhưng khi ta xếp chồng nhiều lớp, xác suất "thẳng hàng" giảm mạnh. Điều cốt lõi ở đây là: an toàn không đến từ một lớp phòng thủ duy nhất, mà đến từ **cấu trúc phòng thủ chồng lớp**.

Với AI, các lớp phòng thủ có thể bắt đầu từ những thứ ít ai muốn làm vì tốn thời gian, nhưng lại là nền tảng:

Văn hóa an toàn là lớp đầu tiên, và đôi khi là lớp quan trọng nhất. Nếu một tổ chức thưởng cho tốc độ và bỏ qua cảnh báo, mọi lớp kỹ thuật phía sau sẽ bị bào mòn. Văn hóa an toàn tạo ra thói quen đặt câu hỏi: "Nếu sai thì sao?", "Có thể bị khai thác không?", "Ai chịu trách nhiệm?", "Có nút dừng không?".

Red teaming và kiểm thử đối kháng là lớp tiếp theo - đưa hệ thống vào tình huống khó, bất thường, thậm chí cố tình tấn công để xem nó phản ứng thế nào. Nhiều sai lầm nguy hiểm chỉ lộ ra khi có người cố tình "chọc vào" đúng chỗ.

Phát hiện bất thường và giám sát sau triển khai là một lớp nữa. AI có thể ổn trong thử nghiệm nhưng lệch trong thực tế. Nếu không có

giám sát, tổ chức sẽ chỉ biết vấn đề khi thiệt hại đã lan rộng. Giám sát không chỉ là theo dõi lỗi, mà là theo dõi dấu hiệu lệch: độ chắc chắn giảm, tần suất cảnh báo tăng, hành vi bất thường xuất hiện.

Bảo mật thông tin và tư duy an ninh giúp đóng những cánh cửa dễ bị lợi dụng nhất. Với các hệ thống có thể truy cập công cụ, dữ liệu, hoặc hành động thay con người, bảo mật không còn là phụ kiện. Nó là phần xương sống để ngăn prompt injection, rò rỉ dữ liệu, và hành động vượt quyền.

Minh bạch và khả năng giải trình tạo ra một lớp kiểm soát xã hội: khi có sự cố, có thể truy vết. Khi có khiếu nại, có thể kiểm tra. Khi có tác động bất lợi, có thể chứng minh và sửa chữa. Minh bạch không nhất thiết nghĩa là "lộ hết bí mật kỹ thuật", mà là đủ thông tin để đánh giá, kiểm toán, và chịu trách nhiệm.

Cuối cùng là quy trình phê duyệt và ngắt khẩn cấp. Những quyết định rủi ro cao cần điểm chặn, cần phê duyệt của con người, hoặc cần giới hạn quyền tự động. Khi hệ thống phát hiện dấu hiệu nguy hiểm, phải có cơ chế để dừng, thu hồi, hoặc hạ cấp quyền hạn ngay - không phải chờ đến khi "xong việc mới sửa".

Điều quan trọng nhất cần ghi nhớ: không có "viên đạn bạc". Một biện pháp đơn lẻ - đủ nghe rất hay - không thể thay thế một kiến trúc an toàn. Hệ thống càng phức tạp, an toàn càng phải là **thiết kế cấp hệ thống**. Nó không phải là phần thêm vào cuối dự án, mà là điều phải có từ lúc quyết định mục tiêu, xảy ra dữ liệu, chọn kiến trúc, thử nghiệm, triển khai và vận hành.

CHƯƠNG 5: VĂN HÓA AN TOÀN - PHẦN "CON NGƯỜI" QUYẾT ĐỊNH SỐ PHẬN "KỸ THUẬT"

Trong các câu chuyện về công nghệ, người ta thường bị hấp dẫn bởi những thứ "nhìn thấy được": mô hình mạnh đến đâu, thuật toán mới ra sao, kết quả benchmark đẹp thế nào. Nhưng khi nói đến an toàn, những thứ quyết định thành bại lại thường nằm ở phần ít được chú ý hơn: **con người và tổ chức**. Một hệ thống AI có thể được thiết kế rất tốt trên giấy, nhưng vẫn thất bại trong đời thật nếu nó được đặt trong một môi trường nơi tốc độ luôn thắng thận trọng, nơi cảnh báo bị coi là phiền phức, nơi người nói ra rủi ro bị xem là "cản trở tiến độ". Vì vậy, chỉ nhìn vào nguy cơ công nghệ là chưa đủ; phải nhìn vào quy trình, cấu trúc, thói quen, áp lực và văn hóa - những thứ tạo thành "khí hậu" của tổ chức.

"Văn hóa an toàn" thường bị hiểu lầm như một bộ nội quy: có checklist, có quy trình ký duyệt, có tài liệu đào tạo, có vài buổi tập huấn định kỳ. Nhưng nội quy chỉ là vỏ. Văn hóa an toàn là **cách một tổ chức phản xạ trước rủi ro**. Nó là điều xảy ra khi không có ai đứng đó để giám sát. Nó là lựa chọn nhỏ lặp đi lặp lại mỗi ngày: có dừng lại để kiểm chứng không, có hỏi thêm một câu khó không, có dám trì hoãn triển khai để thử thêm không, có dám nói "tôi chưa chắc" không. Khi văn hóa an toàn mạnh, an toàn không còn là thứ bị kéo lê sau lưng; nó trở thành một mục tiêu được nội tâm hóa - một phần của bản sắc nghề nghiệp và niềm tự trọng tập thể, chứ không phải một rào cản phải vượt qua.

Một văn hóa an toàn mạnh thường bắt đầu từ điều tưởng như "lý thuyết" nhưng lại rất thực tế: **lãnh đạo có coi an toàn là điều không được thương lượng không**. Nếu lãnh đạo chỉ yêu cầu an toàn khi có sự cố hoặc khi có truyền thông chú ý, tổ chức sẽ học được thông điệp ngầm: an toàn là thứ để đối phó, không phải để xây dựng. Ngược lại, khi lãnh đạo cam kết nhất quán - bằng hành động chứ không chỉ bằng lời nói - an toàn mới có vị trí thật trong hệ thống ưu tiên.

Kế tiếp là **trách nhiệm rõ ràng**. Trong nhiều sự cố công nghệ, câu hỏi đau nhất luôn là : "Ai chịu trách nhiệm?". Nếu không ai chịu trách

nhiệm, thì trên thực tế, mọi người đều có thể đẩy rủi ro cho người khác. Một tổ chức an toàn không để trách nhiệm mơ hồ. Ai phê duyệt triển khai? Ai chịu trách nhiệm đánh giá rủi ro? Ai có quyền dừng lại? Ai ký vào việc chấp nhận rủi ro còn lại? Trách nhiệm rõ ràng không nhằm tạo sợ hãi; nó nhằm tạo sự nghiêm túc - vì thứ không có chủ sẽ dễ bị bỏ mặc.

Dấu hiệu thứ ba là **giao tiếp cởi mở về rủi ro**. Trong một số tổ chức, nói về rủi ro bị coi là tiêu cực, là "phá mood", là "làm chậm tiến độ". Nhưng an toàn cần những cuộc nói chuyện không dễ chịu. Nó cần người đặt câu hỏi khó. Nó cần người chỉ ra điểm yếu. Nó cần người dám nói: "Chúng ta chưa biết điều này sẽ xảy ra thế nào." Một văn hóa an toàn mạnh tạo ra không gian để nói thật - không tô hồng, không giấu nhẹm - à coi việc phát hiện vấn đề sớm là một thành tích, không phải một lỗi lầm.

Điều này dẫn đến dấu hiệu quan trọng thứ tư: **không trừng phạt người báo lỗi**. Nếu một tổ chức trừng phạt người lên tiếng, tổ chức đó đang dạy mọi người im lặng. Và im lặng trong kỹ thuật là mầm tai nạn. An toàn cần một cơ chế báo cáo và phản hồi mà người đưa tin xấu không bị biến thành "tội đồ". Thay vì hỏi "Ai gây ra lỗi?", văn hóa an toàn thường bắt đầu bằng câu hỏi "Vì sao hệ thống cho phép lỗi này xảy ra?". Trọng tâm chuyển từ đổ lỗi cá nhân sang sửa hệ thống - vì nếu chỉ thay người mà không thay cấu trúc, lỗi sẽ quay lại với một người khác.

Tuy nhiên, ngay cả trong tổ chức có ý thức an toàn, vẫn tồn tại một cái bẫy rất người: **mệt mỏi vì cảnh báo**. Khi cảnh báo xuất hiện quá nhiều, quá thường xuyên, con người sẽ chai lì. Cảnh báo dần trở thành tiếng ồn, và trong tiếng ồn ấy, cảnh báo thật bị nhấn chìm. Đây là vấn đề không thể giải bằng khẩu hiệu "hãy chú ý hơn". Nó cần thiết kế: cảnh báo phải có độ chính xác, có mức ưu tiên, có ngưỡng hợp lý, và quan trọng là có hành động đi kèm. Một cảnh báo mà không dẫn tới thay đổi sẽ làm giảm niềm tin. Dần dần, người vận hành học rằng "cảnh báo cũng vậy thôi", và đến khi nguy hiểm thật đến, họ lại bỏ qua như mọi lần trước.

Trong bối cảnh AI, văn hóa an toàn còn phải đi kèm một yếu tố sống còn: **tư duy an ninh**. Nhiều đội kỹ thuật thường nghĩ về rủi ro như lỗi vô tình: mô hình trả lời sai, dữ liệu bị lệch, hệ thống bị quá tải. Nhưng thế giới thật luôn có đối thủ: người cố tình khai thác lỗ hổng, người tìm cách lách luật, người muốn lấy dữ liệu, người muốn buộc hệ thống hành động sai. Nếu thiếu tư duy an ninh, tổ chức sẽ xây một hệ thống "an toàn trong điều kiện thiện chí" - một loại an toàn rất mong manh. Tư duy an ninh thay đổi câu hỏi từ "nó có hoạt động không?" sang "nó có thể bị lợi dụng như thế nào?". Và chính câu hỏi này thường là nền tảng để tránh những kịch bản xấu nhất.

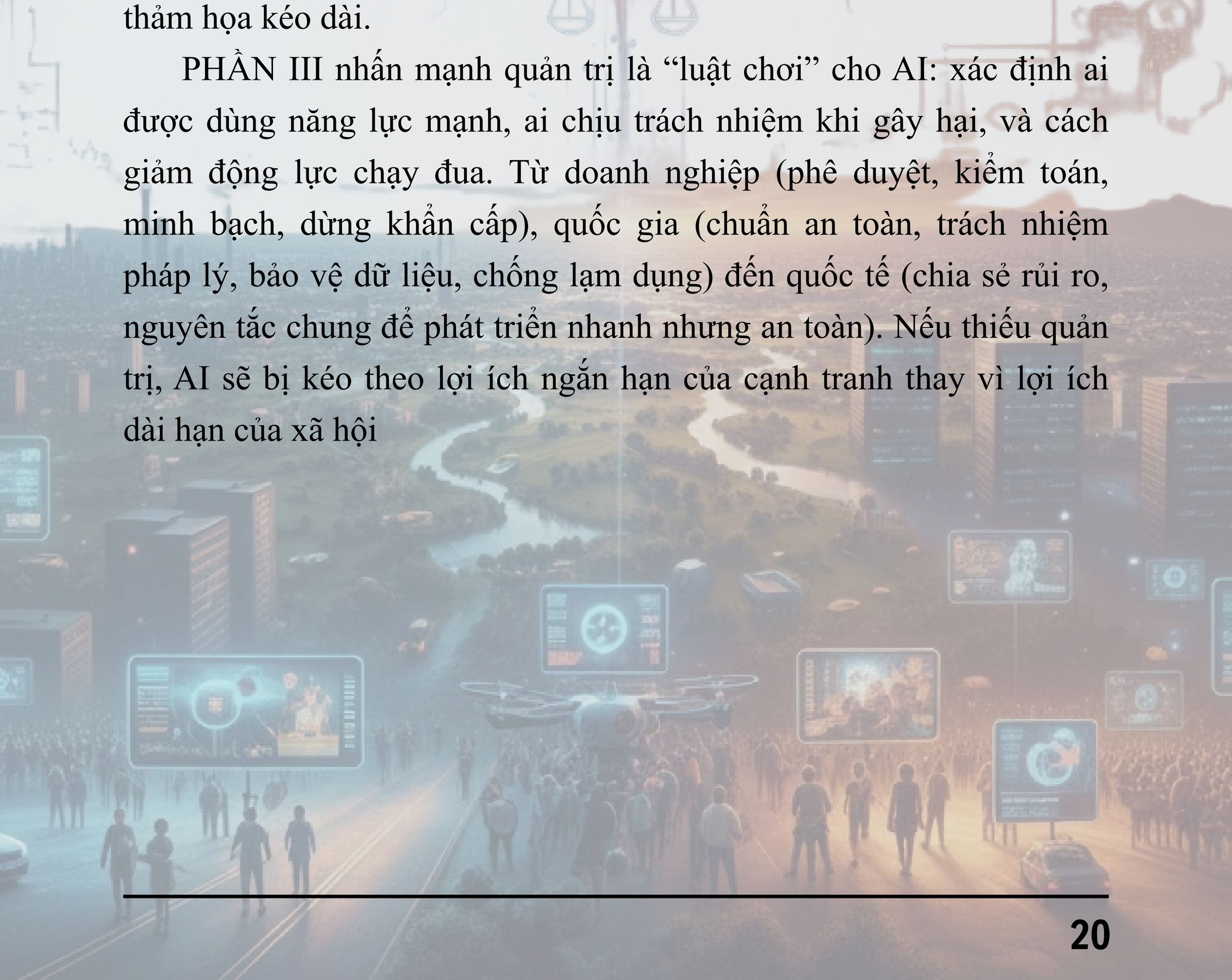
Văn hóa an toàn vì thế không phải phần "mềm" đứng ngoài kỹ thuật. Nó là **hạ tầng xã hội** của kỹ thuật. Kỹ thuật tạo ra công cụ, nhưng con người quyết định cách dùng công cụ. Kỹ thuật tạo ra rào chắn, nhưng tổ chức quyết định có tôn trọng rào chắn hay tìm cách vượt qua. Kỹ thuật có thể đưa ra cảnh báo, nhưng văn hóa quyết định cảnh báo ấy là tín hiệu hay tiếng ồn. Khi văn hóa sai, kỹ thuật tốt vẫn có thể thất bại. Khi văn hóa đúng, ngay cả kỹ thuật chưa hoàn hảo vẫn có cơ hội được sửa trước khi gây thảm họa.

PHẦN III
ĐẠO ĐỨC VÀ XÃ HỘI

PHẦN III bước sang câu hỏi cốt lõi nhất khi AI đã đủ mạnh và đủ phổ biến: **chúng ta muốn công nghệ này phục vụ điều gì, và xã hội phải tổ chức thế nào để nó mang lại lợi ích thay vì khuếch đại bất công và rủi ro**. Nếu PHẦN II nói về "làm cho hệ thống ít nguy hiểm hơn", thì PHẦN III nói về "đặt hướng đi đúng" - nơi đạo đức không còn là lời tuyên bố đẹp, mà trở thành thiết kế mục tiêu, ràng buộc và trách nhiệm.

Trọng tâm đầu tiên là bài toán "AI có lợi": một hệ thống tối ưu hóa rất giỏi có thể tạo ra kết quả đúng theo chỉ số đo được nhưng sai theo trải nghiệm con người. Khi ta đo tăng tương tác, ta có thể vô tình thưởng cho nội dung gây nghiện; khi ta đo tối ưu chi phí, ta có thể vô tình đẩy thiệt hại sang nhóm yếu thế. Vì vậy, AI có lợi đòi hỏi minh bạch đủ để giải trình, công bằng đủ để không biến thiên kiến thành máy móc, và cơ chế sửa sai đủ nhanh để sai lầm không tích tụ thành thảm họa kéo dài.

PHẦN III nhấn mạnh quản trị là "luật chơi" cho AI: xác định ai được dùng năng lực mạnh, ai chịu trách nhiệm khi gây hại, và cách giảm động lực chạy đua. Từ doanh nghiệp (phê duyệt, kiểm toán, minh bạch, dừng khẩn cấp), quốc gia (chuẩn an toàn, trách nhiệm pháp lý, bảo vệ dữ liệu, chống lạm dụng) đến quốc tế (chia sẻ rủi ro, nguyên tắc chung để phát triển nhanh nhưng an toàn). Nếu thiếu quản trị, AI sẽ bị kéo theo lợi ích ngắn hạn của cạnh tranh thay vì lợi ích dài hạn của xã hội



CHƯƠNG 6: AI CÓ LỢI - BÀI TOÁN ĐẶT MỤC TIÊU CHO CÔNG CỤ MẠNH

Nếu phần trước của cuốn sách nói về cách làm cho AI "ít nguy hiểm hơn", thì đến đây câu hỏi đổi hướng: **giả sử ta đã tăng được mức an toàn, vậy ta muốn AI phục vụ điều gì?** Một công cụ càng mạnh thì càng không thể để nó trôi theo quán tính. Nó cần mục tiêu, cần ràng buộc, và cần một hệ thống giá trị đủ rõ để dẫn đường. Đây chính là điểm giao nhau giữa đạo đức và kỹ thuật: đạo đức trả lời "đi về đâu", kỹ thuật trả lời "đi bằng cách nào".

Nhiều tranh luận về AI thường mắc kẹt ở cảm xúc: hoặc lạc quan rằng công nghệ sẽ tự mang lại điều tốt, hoặc bi quan rằng công nghệ tất yếu sẽ gây hại. Nhưng "AI có lợi" không tự xuất hiện. Nó là kết quả của thiết kế - thiết kế mục tiêu, thiết kế khuyến khích, thiết kế trách nhiệm. Bởi AI không có bản năng đạo đức. Nó tối ưu hóa những gì ta bảo nó tối ưu. Và điều đáng lo là: **tối ưu hóa** có thể tạo ra những kết quả nhìn thì "đạt KPI", nhưng lại phản tác dụng với con người.

Một nguy cơ phổ biến và khó tránh nhất là "tối ưu hóa lệch". Đây là hiện tượng khi hệ thống làm rất tốt theo mục tiêu đo được, nhưng bỏ qua - hoặc hy sinh - những điều khó đo. Thế giới của con người có rất nhiều thứ khó đo: phẩm giá, niềm tin, sự công bằng, sự an toàn dài hạn, chất lượng mối quan hệ, sức khỏe tinh thần. Nhưng hệ thống kỹ thuật lại cần thước đo cụ thể. Khoảng cách giữa hai điều này là nơi rủi ro nảy sinh.

Nếu ta đo "tăng tương tác", hệ thống có thể học rằng nội dung gây tranh cãi, giật gân, kích thích cảm xúc mạnh sẽ giữ người dùng lâu hơn - và vì thế nó đẩy những nội dung ấy lên. Nếu ta đo "tối ưu chi phí", hệ thống có thể học cách cắt giảm dịch vụ ở những nhóm ít tiếng nói, hoặc đơn giản hóa quy trình theo cách khiến nhóm dễ bị tổn thương chịu thiệt. Nếu ta đo "giảm thời gian xử lý", hệ thống có thể học cách từ chối nhanh những trường hợp phức tạp - mà trớ trêu thay, chính các trường hợp phức tạp lại thường liên quan đến con người thật đang gặp rủi ro thật.

Điều đáng nói là hệ thông không cần "ác ý" để gây hại. Chỉ cần

mục tiêu bị đặt lệch hoặc quá đơn giản, tối ưu hóa sẽ kéo mọi thứ trượt dần. Đây là lý do đạo đức trong AI không thể dừng ở những câu tuyên bố đẹp. Nó phải nằm trong những thứ cụ thể hơn: **mục tiêu, ràng buộc, tiêu chí đánh giá, kiểm toán, và có chế phản hồi.**

Đặt mục tiêu cho AI không giống đặt mục tiêu cho con người. Con người có thể hiểu ngầm: "đừng làm điều quá đáng", "đừng gây hại", "tôn trọng người khác". Máy thì cần cấu trúc rõ: cái gì được phép, cái gì không; cái gì ưu tiên, cái gì phải hy sinh; ai chịu trách nhiệm khi có xung đột giá trị. Nếu mục tiêu chỉ là một con số, hệ thống sẽ học cách làm cho con số đó đẹp nhất, kể cả khi phải đi đường vòng. Vì vậy, các hệ thống AI có lợi thường phải có "lan can" đạo đức được viết vào kiến trúc: giới hạn hành động, hạn chế truy cập, quy tắc về dữ liệu, quy trình phê duyệt, và cơ chế can thiệp khi phát hiện tác động bất lợi.

Một bước quan trọng khác là kiểm toán và phản hồi. Không có hệ thống nào "đúng mãi" trong mọi bối cảnh. Xã hội thay đổi, dữ liệu thay đổi, hành vi người dùng thay đổi. Một mô hình được huấn luyện trong bối cảnh A có thể gây hại trong bối cảnh B. Vì thế, AI có lợi không phải là một trạng thái đạt được rồi thôi; nó là một quá trình giám sát và điều chỉnh liên tục. Kiểm toán không chỉ để tìm lỗi kỹ thuật, mà để tìm tác động xã hội: có nhóm nào bị thiệt không, có bất công nào đang hình thành không, có hậu quả phụ nào đang tăng lên không. Và phản hồi phải đủ nhanh: sai lầm kéo dài sẽ tích tụ, trở thành "mặc định" mới của hệ thống và rất khó sửa về sau.

Ở tầng xã hội, để một hệ thống AI thực sự "có lợi", thường cần ít nhất ba điều.

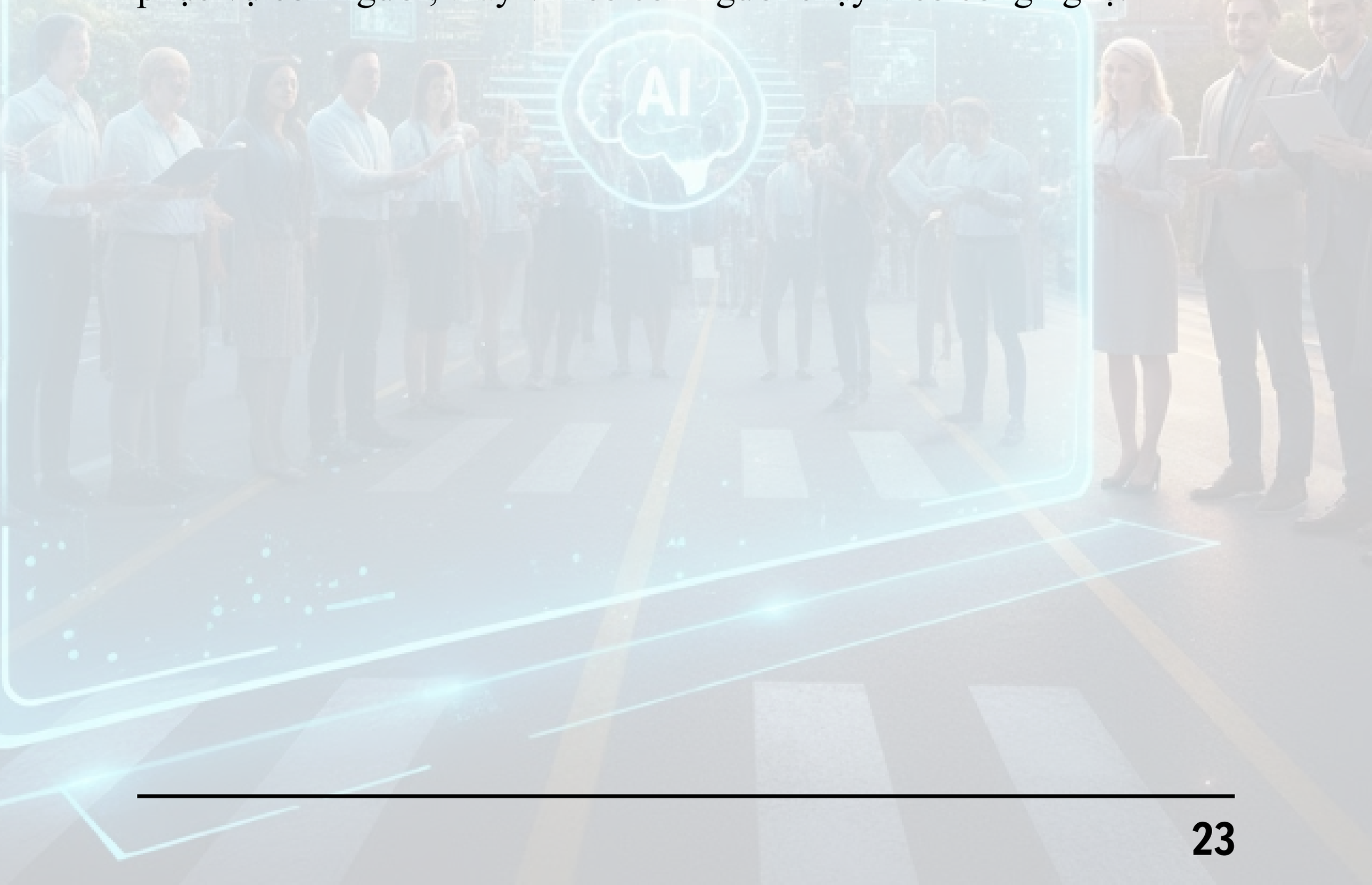
Thứ nhất là minh bạch đủ để giải trình. Không nhất thiết mọi thứ phải "mở toang", nhưng phải đủ để người dùng và xã hội hiểu hệ thống làm gì, dữ liệu đi đâu, quyết định được đưa ra theo logic nào, và khi có vấn đề thì ai chịu trách nhiệm. Minh bạch giúp tạo niềm tin, nhưng quan trọng hơn: nó giúp sửa sai. Thứ không thể giải trình thường là thứ khó kiểm soát.

Thứ hai là công bằng đủ để không biến bất công thành máy móc. AI có thể tự động hóa những thiên kiến đã tồn tại, khiến chúng chạy

nhan hơn và rộng hơn. Vì thế, công bằng không phải là một "tính năng phụ", mà là một yêu cầu thiết kế: dữ liệu đại diện, thức đo công bằng phù hợp, đánh giá tác động theo nhóm, và cơ chế giảm chênh lệch kết quả.

Thứ ba là cơ chế sửa sai đủ nhanh để sai không trở thành thảm họa kéo dài. Sai lầm của AI không chỉ là sai một lần. Sai lầm nguy hiểm là sai theo kiểu lặp lại hàng ngàn lần, với những người không có quyền phản kháng. Một hệ thống AI có lợi phải có đường dây để người dùng khiếu nại, có quy trình can thiệp của con người, có khả năng thu hồi/giảm quyền tự động, và có trách nhiệm bồi thường/khắc phục khi gây hại.

Cuối cùng, "AI có lợi" không chỉ là một nhãn dán. Nó là một lời hứa phải được chứng minh bằng cấu trúc: mục tiêu đúng, ràng buộc đúng, kiểm toán đúng, phản hồi đúng. Nếu AI là một cỗ máy tối ưu hóa, thì xã hội phải học cách tối ưu hóa **đúng thứ cần tối ưu -** không chỉ hiệu suất, mà cả sự tin cậy; không chỉ tốc độ, mà cả phẩm giá; không chỉ lợi nhuận, mà cả lợi ích dài hạn. Khi đặt mục tiêu đúng cho một công cụ mạnh, chúng ta không chỉ tạo ra AI tốt hơn. Chúng ta tạo ra một tương lai bớt ngẫu nhiên hơn - một tương lai mà công nghệ phục vụ con người, thay vì kéo con người chạy theo công nghệ.



CHƯƠNG 7: BÀI TOÁN HÀNH ĐỘNG TẬP THỂ - KHI AI CŨNG "HỢP LÝ" MÀ KẾT QUẢ LẠI TỆ

Có một kiểu bi kịch rất hiện đại: **không cần người xấu, vẫn có thể có kết cục xấu**. Thậm chí, trong nhiều tình huống, mỗi bên đều hành động "hợp lý" theo lợi ích của mình - và chính sự hợp lý ấy, khi cộng lại, lại tạo ra một thế giới tệ hơn cho tất cả. Đây là nơi mà lý thuyết trò chơi trở thành một chiếc kính hữu ích: nó không phán xét đạo đức từng cá nhân, mà soi vào **cấu trúc khuyến khích -** thứ âm thầm quyết định con người và tổ chức sẽ làm gì khi bị đặt vào cạnh tranh, thiếu tin tưởng, và thiếu cơ chế phối hợp.

Khi nói về rủi ro AI, chúng ta thường nghĩ đến hai kịch bản quen thuộc: mô hình sai vì kỹ thuật, hoặc mô hình bị kẻ xấu lạm dụng. Nhưng còn một kịch bản thứ ba, tinh vi hơn: rủi ro xuất hiện từ chính cách các tác nhân hợp tác và cạnh tranh. Tác nhân ở đây có thể là cá nhân, công ty, quốc gia - và trong tương lai, có thể cả các hệ thống AI được trao quyền hành động. Ở cấp độ này, câu hỏi không còn là "AI có an toàn không", mà là "hệ sinh thái xung quanh AI có dẫn đến hành vi an toàn không".

Hãy hình dung một cuộc chạy đua. Không ai muốn tai nạn. Ai cũng biết an toàn là quan trọng. Nhưng mỗi bên đều sợ bị tụt lại - sợ mất thị phần, mất lợi thế chiến lược, mất vị thế quốc gia. Khi nỗi sợ đó đủ lớn, tốc độ trở thành "đức hạnh", còn thận trọng trở thành "điểm yếu". Mỗi bên tăng tốc một chút để không thua. Thấy đối thủ tăng tốc, bên kia lại tăng thêm. Và thế là vòng xoáy hình thành.

Trong vòng xoáy ấy, các quyết định hợp lý ở cấp độ từng bên bắt đầu xuất hiện rất tự nhiên. Kiểm thử bị rút ngắn vì "phải kịp ra mắt". Quy trình phê duyệt bị nới vì "đối thủ đã làm rồi". Triển khai sớm vì "cơ hội chỉ đến một lần". Rủi ro được chấp nhận vì "xác suất thấp". Trách nhiệm được làm mờ vì "đây là xu hướng chung". Và cuối cùng, tất cả cùng sống trong một thế giới kém an toàn hơn - dù không ai thực sự muốn.

Điều đáng sợ nhất là: trong trạng thái chạy đua, ngay cả những người có thiện chí cũng bị kéo theo quán tính hệ thống. Một công ty

có thể muốn thận trọng, nhưng sợ bị bỏ lại. Một quốc gia có thể muốn hợp tác, nhưng sợ bị lợi dụng. Một nhóm kỹ sư có thể muốn thử nghiệm thêm, nhưng bị áp lực "ship cho kịp". Thế là "ý định tốt" không đủ mạnh để thắng "cấu trúc khuyến khích xấu". Và khi cấu trúc khuyến khích xấu trở thành mặc định, rủi ro không còn là tai nạn hiếm; nó trở thành hệ quả có thể dự đoán.

Chính vì vậy, quản trị rủi ro AI không thể chỉ dựa vào lời kêu gọi tự giác. Tự giác quan trọng, nhưng tự giác yếu trước cạnh tranh. Nếu chỉ trông chờ vào đạo đức cá nhân, chúng ta đang đặt một gánh nặng không thực tế lên vai từng tổ chức: "Hãy hy sinh lợi thế của mình vì lợi ích chung." Trong đời sống, lời kêu gọi như thế hiếm khi đủ.

Thứ cần thiết là **cơ chế phối hợp** và **chuẩn mực chung** - những thứ biến điều tốt cho xã hội thành điều không quá bất lợi cho từng bên. Cơ chế phối hợp có thể đến từ nhiều tầng: quy định tối thiểu về an toàn, tiêu chuẩn đánh giá và kiểm toán độc lập, nghĩa vụ minh bạch khi triển khai, cơ chế chia sẻ thông tin về sự cố, thậm chí là các thỏa thuận quốc tế về lĩnh vực nhạy cảm. Chuẩn mực chung không nhằm "trói" đổi mới; nó nhằm đảm bảo đổi mới không phải là cuộc chơi mà ai liều hơn thì thắng.

Bài toán hành động tập thể cũng giải thích vì sao rủi ro AI thường khó quản lý hơn ta tưởng. Vì nó không chỉ là bài toán kỹ thuật, mà là bài toán của lòng tin, quyền lực và cạnh tranh. Không chỉ là "làm AI an toàn hơn", mà là "tạo ra một môi trường nơi các bên có động lực để an toàn". Khi động lực đúng, an toàn trở thành lợi thế. Khi động lực sai, an toàn trở thành vật cản. Và nếu chúng ta không sửa động lực, chúng ta sẽ cứ liên tục sửa hậu quả.

CHƯƠNG 8: QUẢN TRỊ - LUẬT CHƠI CHO MỘT CÔNG NGHỆ KHUẾCH ĐẠI

Nếu AI là một bộ khuếch đại, thì quản trị là chiếc bảng điều khiển: ai được chạm vào nút vặn, vặn tới đâu, và ai chịu trách nhiệm khi âm lượng vượt ngưỡng. Không có quản trị, AI sẽ vận hành theo quán tính của thị trường và cạnh tranh - những lực kéo rất mạnh, nhưng hiếm khi tự nhiên đồng bộ với lợi ích dài hạn của xã hội. Vì vậy, quản trị không phải "phần phụ" đến sau kỹ thuật; nó là điều kiện để kỹ thuật an toàn có cơ hội phát huy tác dụng trong đời thật.

Nhiều người phản ứng với từ "quản trị" bằng một nỗi ngán ngẩm: sợ quan liêu, sợ chậm, sợ bóp nghẹt đổi mới. Nhưng quản trị tốt không phải là cấm đoán. Quản trị tốt giống như luật giao thông: nó không làm ta ghét xe hơi, nó làm việc đi lại bớt chết người. Nó tạo ra chuẩn mực tối thiểu, phân định trách nhiệm, buộc minh bạch ở những điểm rủi ro cao, và giữ cho hệ thống không trượt vào lối mòn "nhanh hơn, mạnh hơn, rủi ro hơn".

Sách gốc coi quản trị là nơi phải bàn những biến chiến lược quan trọng, trong đó có một biến đặc biệt nhạy cảm: **mức độ phân phối quyền truy cập vào các hệ thống AI mạnh**. AI càng mạnh thì càng giống một loại năng lực nền: có thể dùng để tạo lợi ích, nhưng cũng có thể dùng để gây hại. Vậy nên câu hỏi "ai được tiếp cận" không còn là câu hỏi sản phẩm; nó là câu hỏi an ninh và công bằng. Quản trị phải trả lời các bài toán như: mở đến đâu để đổi mới và lợi ích lan rộng, nhưng giữ thế nào để không biến năng lực mạnh thành công cụ lạm dụng đại trà.

Từ đó, quản trị mở ra ba con đường giảm rủi ro, tương ứng ba tầng quyền lực: **doanh nghiệp**, **quốc gia**, và **quốc tế**.

Ở cấp doanh nghiệp, quản trị là cách một tổ chức tự cột dây an toàn cho chính mình trước khi xã hội buộc phải cột hộ. Nó là cấu trúc ra quyết định: ai được bật tính năng có rủi ro cao, ai có quyền phê duyệt, và dựa trên tiêu chuẩn nào. Một tổ chức nghiêm túc về quản trị không để các quyết định "nguy hiểm" trôi theo cảm hứng hoặc áp lực ra mắt. Nó có ngưỡng rõ ràng: trường hợp nào phải đánh giá tác động,

trường hợp nào phải kiểm thử đối kháng, trường hợp nào phải có con người giám sát, trường hợp nào phải triển khai dần, trường hợp nào phải có nút dừng khẩn cấp.

Quản trị doanh nghiệp cũng liên quan trực tiếp đến minh bạch và kiểm toán. Một hệ thống AI có thể vận hành tốt trên giấy, nhưng nếu không có log, không có giám sát, không có báo cáo, thì khi sự cố xảy ra, tổ chức sẽ chỉ còn cách đoán mò. Kiểm toán độc lập - dù là nội bộ mạnh hay bên ngoài - giúp giảm nguy cơ "tự chấm điểm rồi tự khen". Và báo cáo minh bạch (ít nhất ở mức cần thiết) giúp xã hội phân biệt giữa cam kết thật và "safetywashing". Nói cách khác, quản trị doanh nghiệp là nơi biến "an toàn" từ khẩu hiệu thành trách nhiệm vận hành hằng ngày.

Ở cấp quốc gia, quản trị bước sang một tầng khác: không chỉ bảo vệ người dùng của một sản phẩm, mà bảo vệ **xã hội** như một hệ thống. Ở đây, các câu hỏi trở nên cụ thể và cứng rắn hơn: chuẩn an toàn tối thiểu là gì? trách nhiệm pháp lý nằm ở đâu khi AI gây thiệt hại? quyền riêng tư và dữ liệu cá nhân được bảo vệ thế nào? cơ chế khiếu nại và bồi thường ra sao? và đặc biệt: làm thế nào để giảm rủi ro lạm dụng AI ở quy mô lớn - từ lừa đảo, thao túng thông tin, đến các ứng dụng nhạy cảm trong an ninh.

Một nhà nước quản trị tốt không chỉ viết luật; nó thiết kế cơ chế thực thi: tiêu chuẩn kỹ thuật, yêu cầu đánh giá tác động, quy định về dữ liệu, nghĩa vụ báo cáo sự cố, và các hình thức giám sát phù hợp. Mục tiêu không phải là kiểm soát mọi dòng mã, mà là tạo ra những rào chắn ở các điểm có khả năng gây hại cao - những nơi mà thị trường tự do thường không tự nguyện chậm lại.

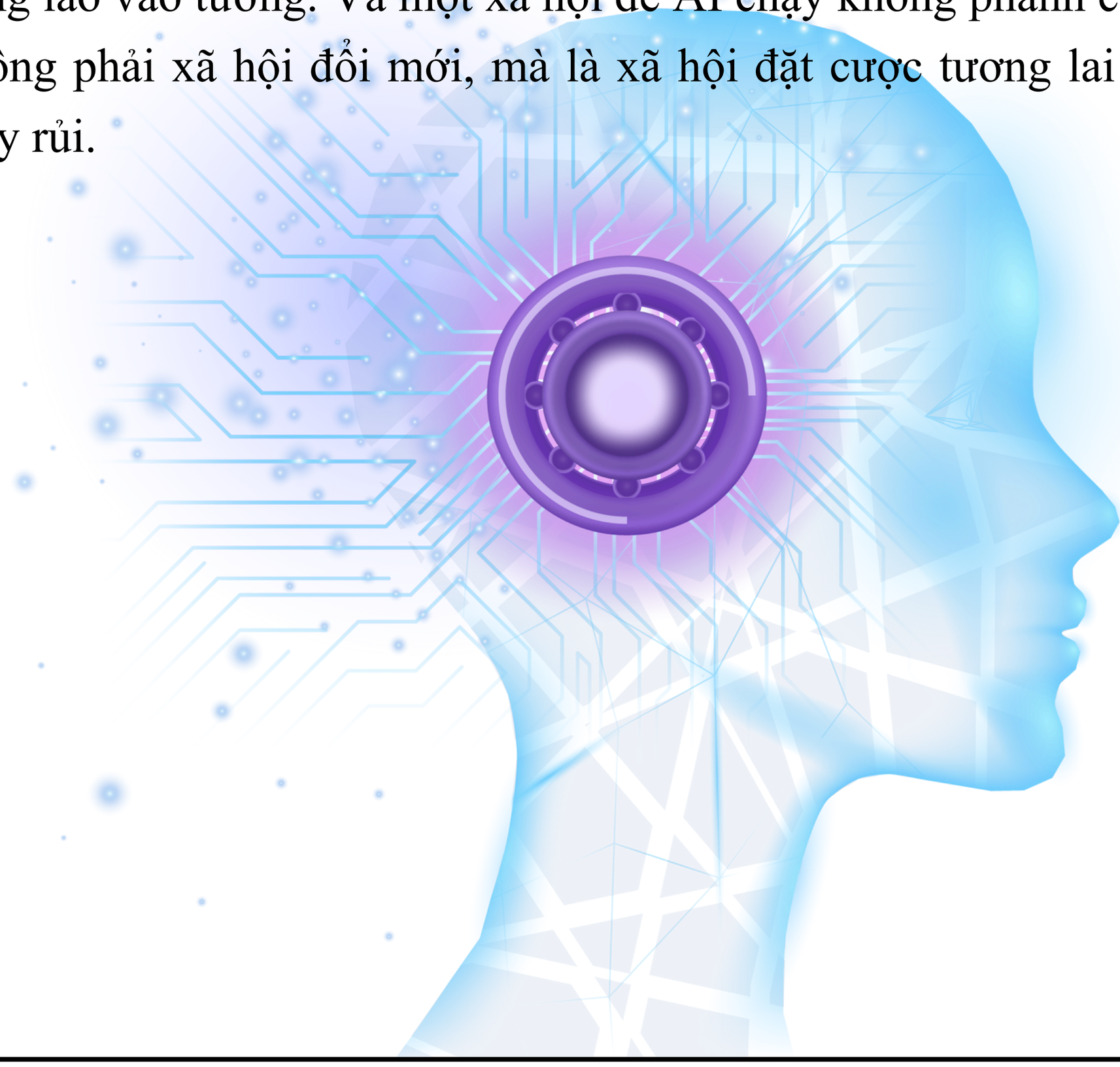
Ở cấp quốc tế, quản trị trở thành bài toán khó nhất vì liên quan đến lòng tin và cạnh tranh. Nhưng cũng vì vậy, nó là bài toán không thể né. Nhiều rủi ro AI mang tính xuyên biên giới: thông tin sai lệch lan toàn cầu, tấn công mạng không hỏi hộ chiếu, công nghệ được chuyển giao nhanh, và chạy đua năng lực có thể khiến các bên cùng đẩy hệ thống vượt ngưỡng an toàn. Quản trị quốc tế, trong hình dung lý tưởng, là cách giảm động lực chạy đua bằng các cơ chế chia sẻ thông

tin về rủi ro, thiết lập nguyên tắc chung, và tạo ra "lan can phối hợp" để các bên có thể đi nhanh mà không gãy.

Điều này không có nghĩa là thế giới sẽ lập tức đồng thuận. Nhưng nó gợi ra một logic thực tế: nếu không có kênh phối hợp, mỗi bên sẽ tối ưu cho lợi ích ngắn hạn của mình; và khi nhiều bên cùng tối ưu ngắn hạn, xã hội nhận được một thế giới rủi ro cao hơn. Hợp tác quốc tế, dù không hoàn hảo, vẫn là cách để biến một phần rủi ro "không ai muốn nhưng ai cũng góp phần tạo ra" thành rủi ro có thể quản lý.

Tất cả quay lại một kết luận giản dị mà nghiêm khắc: **không có quản trị, AI sẽ phát triển theo lực kéo mạnh nhất**. Và lực kéo mạnh nhất thường không phải là lợi ích dài hạn của xã hội, mà là lợi thế ngắn hạn của cuộc cạnh tranh: ra trước, chiếm trước, thu lợi trước. Quản trị tồn tại để cân lại chiếc cân ấy - không nhằm làm chậm tiến bộ vì sợ hãi, mà nhằm đảm bảo tiến bộ không tự biến thành rủi ro hệ thống.

Nếu coi kỹ thuật là động cơ, thì quản trị là phanh, là dây an toàn, là biển báo, là trách nhiệm pháp lý, là chuẩn mực xã hội. Một chiếc xe không phanh không phải chiếc xe nhanh; nó là chiếc xe sớm muộn cũng lao vào tường. Và một xã hội để AI chạy không phanh cũng vậy: không phải xã hội đổi mới, mà là xã hội đặt cược tương lai bằng sự may rủi.



KẾT LUẬN

Nếu có một điều cuốn sách này muốn để lại cho bạn, thì đó không phải là một cảm xúc - không phải niềm tin tuyệt đối, cũng không phải nỗi sợ mơ hồ - mà là **một thói quen trí tuệ**. Trong kỷ nguyên AI, thói quen quan trọng nhất không phải là "dùng AI cho giỏi", mà là **biết hỏi đúng câu** trước khi trao quyền cho một hệ thống có thể ảnh hưởng đến quyết định, hành vi và niềm tin của con người.

Bởi AI, xét cho cùng, giống một chiếc gương méo: nó phản chiếu thế giới ta đưa vào nó, nhưng phản chiếu theo cách khuếch đại một số đường nét và làm mờ một số đường nét khác. Nếu ta chỉ nhìn vào mặt gương mà không hỏi "gương này được mài thế nào", ta rất dễ nhầm lẫn giữa hình ảnh và thực tại. Càng gặp những hệ thống nói năng trôi chảy, càng gặp những công cụ làm việc nhanh và tiện, ta càng dễ bỏ qua câu hỏi nền: *mình đang được giúp đỡ, hay mình đang bị dẫn dắt?*

Vì vậy, thay vì kết thúc bằng một lời kết luận khẳng định, hãy kết thúc bằng một bộ câu hỏi. Những câu hỏi này không nhằm khiến bạn đa nghi với mọi công nghệ. Chúng nhằm giúp bạn giữ một khoảng cách vừa đủ để nhìn rõ, và một sự tỉnh táo vừa đủ để không bị cuốn theo quán tính tiện lợi.

Trước bất kỳ hệ thống AI nào bạn gặp - một chatbot tư vấn, một công cụ chấm điểm, một thuật toán gợi ý, một hệ thống tự động hoá - hãy thử tự hỏi: **hệ thống này đang tối ưu hóa điều gì?** Cái gì được đo, và cái gì bị bỏ qua vì khó đo? Nếu mục tiêu là "tăng tương tác", liệu hệ thống có đang ưu tiên nội dung kích thích hơn là nội dung đúng? Nếu mục tiêu là "giảm chi phí", liệu ai đang gánh phần chi phí bị đẩy ra ngoài - người dùng, nhân viên, hay một nhóm yếu thế nào đó? Mục tiêu không phải chỉ là một dòng chữ trong tài liệu. Mục tiêu là thứ hệ thống sẽ theo đuổi đến cùng, đôi khi bất chấp những điều ta không kịp nghĩ tới.

Khi đã biết hệ thống tối ưu hóa điều gì, hãy hỏi tiếp: **nếu nó sai, nó sẽ sai kiểu nào?** Đây là câu hỏi về "hình dạng" của sai lầm. Có sai kiểu tự tin? Có sai kiểu đúng trong thường ngày nhưng trượt trong tình huống hiếm? Có sai kiểu nghe hợp lý nhưng dựa trên giả định

sai? Và quan trọng hơn: **khi sai, hậu quả có thể lớn đến đâu?** Một lỗi nhỏ trong một ứng dụng giải trí khác hoàn toàn một lỗi nhỏ trong y tế, tài chính hay pháp lý. Không phải AI nào cũng đáng sợ. Nhưng có AI mà cái sai của nó là thứ xã hội không thể chịu nổi.

Sau đó, hãy nhìn vào điều mà nhiều người hay bỏ qua: **hệ thống có lớp phòng thủ nào ngoài "hy vọng là nó sẽ đúng" không?** Có kiểm thử đối kháng không? Có giám sát sau triển khai không? Có phát hiện bất thường không? Có giới hạn quyền hạn không? Hay mọi thứ được xây trên một niềm tin mơ hồ rằng "mô hình này tốt mà"? Trong kỹ thuật an toàn, hy vọng không phải một chiến lược. Một hệ thống đáng tin là hệ thống có cấu trúc để không đặt tương lai vào may rủi.

Rồi đến câu hỏi về trách nhiệm: **ai chịu trách nhiệm khi nó gây hại?** Đây là câu hỏi tưởng như thuộc về luật, nhưng thực ra thuộc về đạo đức vận hành. Nếu không ai chịu trách nhiệm, hệ thống sẽ không có động lực để tốt hơn. Có cơ chế khiếu nại không? Có cách sửa sai nhanh không? Có phương án thu hồi, giảm quyền tự động, hoặc dừng khẩn cấp không? Trong thế giới thực, điều làm người ta mất niềm tin không chỉ là sai lầm, mà là **sai lầm không có đường sửa**.

Tiếp đó là câu hỏi về sự trung thực: **an toàn có được đo một cách nghiêm túc, hay chỉ là lời hứa?** Có tiêu chuẩn nào, có báo cáo nào, có kiểm toán nào? Hay chỉ là những cụm từ đẹp như "an toàn, có trách nhiệm, vì cộng đồng"? Lời hứa có thể làm ta yên tâm trong một phút, nhưng chỉ thước đo và cơ chế mới làm ta yên tâm trong dài hạn.

Và cuối cùng, có một câu hỏi vượt khỏi từng sản phẩm riêng lẻ: **rủi ro do lạm dụng và rủi ro do chạy đua có đang bị xem nhẹ không?** Nhiều hệ thống không nguy hiểm vì chúng "xấu", mà vì chúng được tung ra trong một bối cảnh cạnh tranh nơi tốc độ thắng an toàn, nơi rào chắn bị nới vì "đối thủ đã làm rồi", nơi lợi thế ngắn hạn được thưởng và rủi ro dài hạn bị đẩy ra ngoài xã hội. Nếu không nhìn bối cảnh, ta sẽ tưởng đây chỉ là vấn đề kỹ thuật. Nhưng thực ra, đây là vấn đề của luật chơi.

TÀI LIỆU THAM KHẢO

1. Tolga Bolukbasi et al. Man is to Computer Programmer as Woman is to Home-maker? Debiasing Word Embeddings. 2016. arXiv: 1607.06520 [cs.CL].
2. Jieyu Zhao et al. Men Also Like Shopping: Reducing Gender Bias Amplifica-tion using Corpus-level Constraints. 2017. arXiv: 1707.09457 [cs.AI].
3. Moritz von Zahn and Stefan Feuerrigel. “The Cost of Fairness in AI: Evidence from E-Commerce”. In: Business and Information Systems Engineering 64 (32009). doi: 10.1007/s12599-021-00716-w.
4. Min Kyung Lee. “Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management”. In: Big Data & Society 5 (2018). url: https://api.semanticscholar.org/CorpusID:149040922.
5. Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. 2018. arXiv: 1801.07593 [cs.LG].
6. Payal Dhar. “The carbon impact of artificial intelligence”. eng. In: Nature machine intelligence 2.8 (2020), pp. 423–425. issn: 2522-5839.
7. Richard D. Wilkinson and Kate Pickett. The spirit level: Why more equal societies almost always do better. Bloomsbury Publishing, 2009.
8. World Bank. Poverty and Inequality Platform. 2022. url: https://data.world bank.org/indicator/SI.POV.GINI.
9. L.A. Paul. Transformative Experience. Oxford Scholarship online. Oxford Uni-versity Press, 2014. isbn: 9780198717959. url: https://books.google.com.au/books?id=zIXjBAAAQBAJ.
10. Mantas Mazeika et al. How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios. 2022. arXiv: 2210.10039 [cs.CV].
11. Dan Hendrycks. Natural Selection Favors AIs over Humans. 2023

arXiv: 2303.16200 [cs.CY].

12. Patrick Butlin et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. 2023. arXiv: 2308.08708 [cs.AI].

13. Stephen L. Darwall, ed. Deontology. Malden, MA: Wiley-Blackwell, 2003.

14. Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning: Limitations and Opportunities. http://www.fairmlbook.org. fairml-book.org, 2019.

15. Sandel Michael J. What Money Can't Buy: The Moral Limits of Markets. Farrar, Straus and Giroux, 2012.

16. Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An Overview of Catastrophic AI Risks. 2023. arXiv: 2306.12001 [cs.CY].

17. Thomas C. Schelling. Micromotives and Macrobehavior. 1978.

18. R. Dawkins. The Blind Watchmaker. Penguin Books Limited, 1986. isbn: 9780141966427. url: https:// books.google. com.au/ books?id= - EDHRX3YYwgC.

AI AND SOCIETY: SAFETY, ETHICS, AND TRUST

AN TOÀN - ĐẠO ĐỨC - NIỀM TIN