MỤC LỤC

1

MỞ CHIẾC HỘP ĐEN AI

Trí tuệ nhân tạo (AI) thường bị thần bí hóa và hiểu sai, đặc biệt là trong vấn đề ai là người được hưởng lợi và ai là người bị loại trừ khi nó được ứng dụng. Các nhà công nghệ theo chủ nghĩa lí tưởng thường xem AI như một cuộc cách mạng nhân hậu và vĩ đại, có thể giải quyết mọi vấn đề tưởng chừng không thể – từ khủng hoảng y tế, biến đổi khí hậu, an ninh mạng, thúc đẩy học tập, cho đến bảo vệ đa dạng sinh học. Ngược lại, những người giữ thái độ bi quan với công nghệ lại coi AI là điềm báo rợn người cho những vấn đề xã hội nghiêm trọng, có thể dẫn đến sự diệt vong của nhân loại – nơi mà sự giám sát và kiểm soát từ trên xuống của nhà nước hoặc các tập đoàn, vũ khí tận thế quân sự, và các vấn đề đạo đức sinh học liên quan đến việc tạo ra những đứa trẻ "thiết kế sẵn" bị đẩy đến mức mất kiểm soát. Chủ nghĩa định đoạt công nghệ này – niềm tin rằng công nghệ quyết định cách xã hội tiếp nhận và thay đổi theo sự phát triển của công nghệ mới, dù theo hướng tốt đẹp hay tiêu cực – đang làm tổn hại đến diễn ngôn về việc chúng ta thực sự có thể

định hình và lựa chọn cách sử dụng các hệ thống AI như thế nào. Trong cuốn sách này, tôi sẽ không đứng hẳn về một trong hai thái cực "tốt" hay "xấu" của AI. Thay vào đó, tôi tiếp cận vấn đề một cách thực tế hơn: AI mang lại điều gì cho chúng ta – những người trực tiếp sử dụng? Tại sao các nhà phát triển lại thiết kế AI theo cách này mà không phải cách khác? Và toàn xã hội bị ảnh hưởng ra sao? Qua đó, chúng ta có thể thấy rõ ai đang được hưởng lợi từ AI và ai thì không.

Trí tuệ nhân tạo đang định hình cuộc sống của chúng ta theo vô vàn cách khác nhau. Từ cách chúng ta làm việc, chất lượng dịch vụ y tế mà chúng ta nhận được, đến việc ai đủ điều kiện vay vốn ngân hàng hoặc mua bảo hiểm, hay ai sẽ được ghép đôi trên mạng xã hội, thậm chí cả những món hàng mà chúng ta mua, tất cả đều đang chịu ảnh hưởng từ cụm từ có vẻ bí ẩn: **"AI"**. Nhưng rốt cuộc AI là gì? Và nó tác động đến cuộc sống chúng ta theo *những cách nào*? Điều gì tạo ra thiên kiến ngầm trong hệ thống khiến nó ưu tiên người A hơn người B khi đưa ra quyết định, chẳng hạn như ai được gọi phỏng vấn xin việc, hoặc ai sẽ được tiếp cận một loại thuốc thử nghiệm? Liệu AI có đưa ra quyết định không đồng nhất, dựa trên các đặc điểm cá nhân khác nhau, từ đó mang lại lợi thế cho một số cá nhân hoặc nhóm người nhất định, đồng thời loại trừ những người khác?

AI thường được xem như một *hộp đen*, tức là khi

những công việc khoa học và kĩ thuật bên trong chúng trở nên vô hình bởi chính sự thành công của

> chúng. Khi một cỗ máy vận hành trơn tru và hiệu quả, [...] người ta chỉ chú ý đến đầu vào và đầu ra, mà không cần quan tâm đến sự phức tạp bên trong. Vì thế, một cách nghịch lí, khoa học và công nghệ càng thành công thì chúng lại càng trở nên khó thấy và khó hiểu.
>
> (Latour, 1999, tr. 304)

Việc "đóng hộp đen" AI đồng nghĩa với việc người dùng phổ thông (và thậm chí cả một số chuyên gia tự xưng) hầu như không có kiến thức về cách thức hoạt động hoặc lí do vì sao AI vận hành theo cách mà nó đang vận hành. Cuốn sách này cố gắng hé mở "hộp đen AI", đặt nó trong bối cảnh xã hội cụ thể để làm rõ: *ai* là người chịu ảnh hưởng từ AI, và theo những cách nào. Trọng tâm là các vấn đề liên quan đến sự đa dạng – những gì AI đang tạo ra, những gì nó chưa giải quyết được – cũng như cách mà AI có thể được sử dụng như một công cụ để tăng cường sự hòa nhập và giảm thiểu thiên kiến. AI sẽ mang thiên kiến khi những người thiết kế nó không nhận thức được các hành vi phân biệt đối xử mà họ đã vô tình tích hợp vào hệ thống. Đồng thời, nếu dữ liệu huấn luyện không phản ánh đúng xã hội mà chúng ta mong muốn hướng tới (hoặc không phù hợp với ngữ cảnh sử dụng thực tế), AI cũng sẽ đưa ra các quyết định sai lệch. Chẳng hạn, Wittkower (2016, tr. 1) cho rằng "để giảm thiểu các tác động phân biệt đối xử từ thiết kế kĩ thuật, cần áp dụng một quan điểm thiết kế mang tính chống phân biệt một cách chủ động". Sự phân biệt đối xử của AI có thể xảy ra ở cả cấp độ nhóm, chẳng hạn như một

hệ thống tuyển dụng của công ty không muốn tuyển phụ nữ vì trước đây công ty này vốn đã không làm điều đó, và ở cấp độ cá nhân – ví dụ như một hệ thống nhắm vào một cá nhân với điểm tín dụng (xã hội) thấp nếu người đó khớp với các tham số liên quan đến, chẳng hạn, gian lận bảo hiểm y tế. Tuy nhiên, chỉ nhận thức thôi là chưa đủ. Chúng ta cần có kiến thức về những biến đổi xã hội và vật chất do AI tạo ra, cũng như các phương pháp thử nghiệm và thực hành thay vì chỉ trích vấn đề. Vì AI ngày càng đóng vai trò quyết định trong đời sống hằng ngày, đã đến lúc cần xem xét lại mối quan hệ giữa con người và máy móc (Dräger & Müller-Eiselt, 2020, tr. 7).

Tuy nhiên, nếu chỉ tập trung vào việc nêu bật các ví dụ về việc AI "trục trặc" hoặc đưa ra những quyết định sai lầm – chẳng hạn như trường hợp chatbot Tay của Microsoft trên Twitter (nay là X), sau vài giờ học từ người dùng đã bắt đầu đăng những dòng tweet phân biệt chủng tộc, ví dụ như phủ nhận vụ Diệt chủng Do Thái (Wolf và cộng sự., 2017) – thì chúng ta sẽ bị mắc kẹt trong một *vòng lặp thông tin khép kín*, nơi AI luôn bị gắn liền với hình ảnh những "cỗ máy xấu xa". Điều quan trọng là chúng ta cần xem xét trách nhiệm đạo đức của các nhà phát triển đối với công nghệ mà họ tạo ra. Đồng thời, cũng cần hiểu rằng AI thật sự có thể mang lại lợi ích cho nhân loại nếu được sử dụng đúng cách. AI có thể trở thành một công cụ nâng cao nhận thức về sự đa dạng, bất công chủng tộc, các vấn đề về giới, quyền của cộng đồng LGBTQ+, cũng như hỗ trợ các nhóm yếu thế hoặc chịu sự phân biệt đối xử. Tất nhiên, điều này đòi hỏi chúng ta phải đồng thời nhận thức rõ các nguy cơ của AI và chủ động giảm thiểu chúng.

Một ví dụ cho thấy AI có thể được sử dụng tích cực là giúp các cấp quản lí và tuyển dụng nhận diện thiên kiến trong quy trình tuyển dụng và thăng chức. Nhiều nghiên cứu đã chỉ ra rằng tên gọi, giới tính hoặc ngoại hình của ứng viên có thể ảnh hưởng đến cách họ được đánh giá, thậm chí trước cả khi họ có cơ hội thể hiện năng lực thật sự. Một ví dụ về thiên kiến đến từ con người là việc đính kèm ảnh chân dung trong CV: điều này có thể là bất hợp pháp, được khuyến nghị, hoặc hoàn toàn ngẫu nhiên, tùy theo từng quốc gia. Nếu AI xét đến các định kiến sẵn có của doanh nghiệp thì có thể tạo nền tảng cho việc tuyển dụng đội ngũ nhân sự đa dạng hơn (Cohen, 2019). Một ví dụ khác là AI có thể vẽ bản đồ rủi ro sức khỏe cho một số nhóm nhất định với năng lực tính toán vượt trội. Chẳng hạn, AI có thể xây dựng các cơ sở dữ liệu lớn về rủi ro y tế thông qua nhận diện hình ảnh, từ đó chẩn đoán nhanh và chính xác để mang lại dịch vụ chăm sóc tốt hơn cho bệnh nhân (McKinney và cộng sự, 2020; Pisano, 2020; Salim và cộng sự, 2020; Wang và cộng sự, 2021). Việc nhận thức được thiên kiến từ cả AI lẫn con người sẽ giúp chúng ta giảm thiểu rủi ro. Thu thập thêm mẫu máu từ những người trên 60 tuổi có thể giúp giảm bớt thiên kiến của AI do thiếu hiểu biết về tình trạng sức khỏe của người cao tuổi. Việc xây dựng thêm các cơ sở dữ liệu bao gồm người da màu có thể giúp công nghệ nhận diện khuôn mặt hoạt động chính xác hơn đối với những người không phải da trắng. Tuy vậy, AI vẫn tiềm ẩn nhiều rủi ro nghiêm trọng liên quan đến vấn đề đa dạng.

Trong cuốn sách này, tôi chọn ra năm đặc điểm cá nhân cụ thể để đi sâu vào các vấn đề liên quan đến AI và sự đa dạng. Tôi sẽ phân tích cách AI xử lí chủ đề giới tính và bản dạng giới (Chương 2), cộng đồng LGBTQ+ (Chương 3), chủng tộc (Chương 4), cơ thể, sức khỏe và quá trình lão hóa (Chương 5), tình trạng kinh tế-xã hội và giai cấp (Chương 6). Tôi chọn những nhóm chủ đề này vì chúng là những ví dụ quan trọng cho thấy AI có thể ảnh hưởng đến đời sống con người dựa trên các đặc điểm cá nhân như thế nào. Việc AI được sử dụng trong những lĩnh vực này phản ánh các vấn đề cấu trúc sâu xa trong xã hội, liên quan đến quyền lực, ngôn ngữ, luật pháp và thị trường lao động. Hiểu được cách AI tác động đến một nhóm cụ thể có thể giúp chúng ta nhận diện những thiên kiến mang tính phân biệt liên quan đến các nhóm khác, đồng thời cho thấy hướng phát triển công nghệ trong tương lai.

Mỗi chương trong sách đều khám phá và viện dẫn các nghiên cứu cũng như dự án thực tế liên quan trực tiếp đến các yếu tố này và AI, nhằm mang đến một bức tranh hiện đại về mối quan hệ giữa AI và sự đa dạng vào năm 2023. Ở phần cuối sách, tôi sử dụng các lí thuyết từ lĩnh vực *nghiên cứu khoa học và công nghệ* (science and technology of studies – STS), *Khoa học công nghệ nữ quyền* (Feminist Technoscience) và *giao cắt* (intersectionality) để giải mã cấu trúc huyền thoại về "con người tiêu chuẩn" – bằng cách chỉ ra rằng AI có thể trở nên phân biệt đối xử nếu được phát triển mà không có sự cân nhắc đầy đủ. Chúng ta cần một nền AI có trách nhiệm, nghiêm túc nhìn nhận các vấn đề về sự đa dạng.

Khi hiểu được AI có thể hoạt động vì lợi ích – hoặc đi ngược lại lợi ích – của các nhóm người khác nhau, và đồng thời nhận thức rõ các rủi ro tiềm ẩn, chúng ta sẽ có thể xây dựng và triển khai những hệ thống AI công bằng, bao trùm và vững chắc hơn.

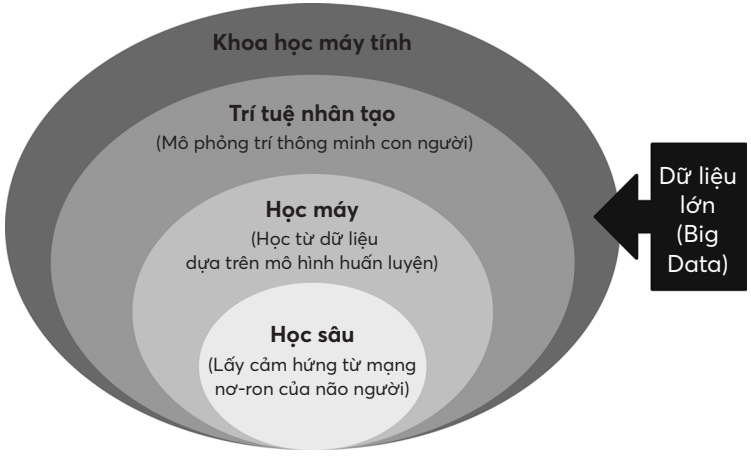
AI LÀ GÌ?

Nếu một người bất kì muốn biết AI là gì, khả năng cao là họ sẽ tra cứu cụm từ này trên mạng hoặc trong từ điển. Tuy nhiên, khi tìm kiếm định nghĩa về AI, kết quả nhận được lại khá đa dạng và không hoàn toàn thống nhất. Từ điển Oxford tiếng Anh (2022) định nghĩa AI là "khả năng của máy tính hoặc các loại máy móc khác trong việc thể hiện hoặc mô phỏng hành vi thông minh". Trong khi đó, Từ điển Merriam-Webster (2022a) lại mô tả AI là "một nhánh của khoa học máy tính liên quan đến việc mô phỏng hành vi thông minh trên máy tính và khả năng của máy móc trong việc bắt chước hành vi thông minh của con người". Các định nghĩa tương tự cũng xuất hiện trong giới nghiên cứu. Ví dụ, Anjila (1984, tr. 65) viết: "Trí tuệ nhân tạo là một nhánh của khoa học và công nghệ chuyên tạo ra các máy móc và chương trình máy tính thông minh để thực hiện nhiều nhiệm vụ khác nhau đòi hỏi trí tuệ con người."

Từ các định nghĩa trên, ta có thể thấy sự khác biệt giữa việc *mô phỏng* trí thông minh giống con người và *sở hữu* trí thông minh giống con người. Bên cạnh đó,

cũng có sự phân biệt giữa việc một cỗ máy có trí thông minh tổng quát hay chỉ có khả năng thực hiện thông minh một số nhiệm vụ cụ thể, giới hạn. AI thường được chia thành hai loại: "AI yếu" (những hệ thống chỉ có thể thực hiện một nhiệm vụ cụ thể một cách có vẻ thông minh) và "AI mạnh" (hệ thống mà theo định nghĩa của con người, có thể tư duy một cách toàn diện, như một trí tuệ thực sự). Một cách gọi khác cho cặp khái niệm này là AI tổng quát và AI hẹp, theo như mô tả của Broussard (2018, tr. 32) rằng "*AI tổng quát* là kiểu AI trong các bộ phim Hollywood... những cỗ máy suy nghĩ như con người. Còn *AI hẹp* thì khác – nó là một phương pháp toán học để đưa ra dự đoán... AI tổng quát là thứ mà một số người mong muốn, còn AI hẹp là thứ chúng ta đang có." Trong cuốn sách này, tôi tập trung chủ yếu vào AI đang hiện hữu trong thực tế, tức AI hẹp.

Thuật ngữ AI lần đầu tiên được đặt ra vào mùa hè năm 1956, khi các nhà khoa học Mỹ gồm John McCarthy, Marvin L. Minsky, Nathaniel Rochester và Claude Shannon tổ chức Hội thảo Nghiên cứu Mùa hè về Trí tuệ nhân tạo tại Dartmouth, New Hampshire (Mỹ). Marvin Minsky sau này đã có nhiều đóng góp quan trọng cho sự phát triển và tư tưởng về AI. Ông từng hợp tác thiết kế nhân vật AI phản diện nổi tiếng HAL 9000 trong bộ phim *2001: A Space Odyssey* của đạo diễn Kubrick (1968). Từ đó đến nay, AI đã trải qua nhiều giai đoạn thăng trầm: khởi đầu đầy lạc quan với sự đầu tư mạnh mẽ về nghiên cứu và phát triển, nhưng cũng từng rơi vào những thời kì trì trệ, được gọi là "mùa đông AI" (giai đoạn 1974-1980 và 1987-1993).



Hình 1.1 AI và các lĩnh vực liên quan

AI là một nhánh của ngành *khoa học máy tính* – lĩnh vực nghiên cứu cả lí thuyết lẫn ứng dụng thực tiễn của máy tính, bao gồm phần cứng, phần mềm và các thuật toán (tức tập hợp các hướng dẫn để giải quyết các vấn đề cụ thể và thực hiện phép tính). Từ "computer" (máy tính) ban đầu được dùng để chỉ những nhân công con người chuyên thực hiện và tính toán các bảng số học (Broussard, 2018, tr. 77). Trong vài thập kỉ gần đây, ngành khoa học máy tính đã phát triển bùng nổ, cả trong giáo dục lẫn thị trường việc làm. AI được xem là một nhánh con trong lĩnh vực khoa học máy tính, và bản thân AI cũng có nhiều phân nhánh nhỏ, được thể hiện trong Hình 1.1. Một trong số đó là *học máy* (machine learning), nơi máy tính học từ dữ liệu, tức là huấn luyện mô hình dựa trên các tập dữ liệu. Thông thường, quá trình học máy diễn ra theo ba cách: học có giám sát (dữ liệu được

gắn nhãn, thường là do con người gắn, giúp hướng dẫn máy học), học không giám sát (máy không được cung cấp nhãn hoặc hướng dẫn cụ thể) và học tăng cường (máy học thông qua "phần thưởng" và "hình phạt" dựa trên phản hồi từ kết quả) (Broussard, 2018, tr. 93). Trong học máy, còn có một nhánh sâu hơn gọi là *học sâu*, là hình thức học máy có cấu trúc mô phỏng các mạng lưới thần kinh phức tạp trong não người. Khoa học máy tính và các nhánh của nó thường tận dụng dữ liệu lớn, tức các tập dữ liệu khổng lồ, cho phép máy tính phát hiện mẫu hình và đưa ra giải pháp dựa trên lượng dữ liệu được cung cấp. Mặc dù cuốn sách này tập trung vào AI, bạn cần lưu ý rằng AI chỉ là một nhánh trong hệ sinh thái rộng lớn của khoa học máy tính, và bản thân nó cũng có nhiều tầng lớp, nhiều thuật ngữ chuyên môn, mà theo thời gian có thể thay đổi. Vì đây không phải là một cuốn sách chuyên về kĩ thuật, nên bạn sẽ không bắt gặp những hướng dẫn lập trình hay các thuật ngữ đòi hỏi trình độ chuyên sâu để hiểu. Tuy vậy, một chút kiến thức kĩ thuật chắc chắn sẽ giúp bạn hiểu rõ hơn về tác động xã hội phức tạp mà các công nghệ này mang lại.

Thuật ngữ "trí tuệ nhân tạo" đã bị chỉ trích vì mô tả một thứ thực ra không hẳn là "nhân tạo" cũng chẳng thật sự "khôn" – như lập luận của Crawford (2021) trong cuốn *Atlas of AI*. Crawford cho rằng AI được tạo nên từ các tài nguyên tự nhiên, và chính con người mới là yếu tố khiến hệ thống trông có vẻ tự động. Ví dụ, Crawford trình bày rằng khi một người đặt mua giấy vệ sinh qua trợ lí giọng nói thông minh, hành động tưởng như đơn giản đó thực chất sẽ kích hoạt cả một chuỗi quy trình phức tạp: từ việc