Trí tuệ nhân tạo: các chuẩn mực kỹ thuật và các quyền cơ bản, một hổn hợp nhiều rủi ro

Tác giả: Mélanie Gornet và Winston Maxwell

Huỳnh Thiện Quốc Việt dịch

Nguồn: Intelligence artificielle: normes techniques et droits fondamentaux, un mélange risqué, The Conversation, ngày 28/09/2022.

Nhãn hiệu “CE” đảm bảo sự tuân thủ các chuẩn mực châu Âu của các sản phẩm được sản xuất. Liệu điều đó có thể thích nghi với trí tuệ nhân tạo hay không? Olga PaHa, Shutterstock

Quy định về “Đạo luật AI” của châu Âu trong tương lai có mục đích là tạo ra một khuôn khổ pháp lý cho toàn bộ các hệ thống trí tuệ nhân tạo (AI), đặc biệt là các hệ thống gây ra những rủi ro quan trọng đối với an ninh hoặc các quyền cơ bản, chẳng hạn như sự không phân biệt đối xử, đời sống riêng tư, quyền tự do ngôn luận hoặc nhân phẩm.

Trong phiên bản hiện tại, bản dự thảo luật xử lý tất cả các rủi ro nói trên theo cùng một cách. Nhãn hiệu CE (tức “tuân thủ chuẩn châu Âu”, giống như nhãn hiệu hiện hành), sẽ chỉ ra rằng hệ thống AI được xem là đủ mức an toàn để được đưa ra thị trường, cho dù đó là an toàn vật lý hay an toàn đối với các quyền cơ bản. Văn kiện được

trình bày dành cho các chuẩn mực kỹ thuật và các phân tích rủi ro một vai trò quan trọng, hy vọng có thể chuẩn mực hóa các phương pháp đánh giá.

Tuy nhiên, vẫn có một sự căng thẳng giữa cách tiếp cận "dựa trên rủi ro", do Ủy ban Châu Âu đề xuất, với cách tiếp cận tập trung vào việc tôn trọng các quyền cơ bản, được các tòa án và các công trình nghiên cứu của Hội đồng châu Âu ủng hộ.

Thật vậy, trong khi dễ dàng hình dung một loạt các thử nghiệm đối với một tiêu chí an toàn, theo cách các thử nghiệm đã được thực hiện trên đồ chơi trẻ em trước khi chúng được đưa ra thị trường, thì việc đánh giá sự không phân biệt đối xử theo cách này sẽ khó hơn [đối với hệ thống AI], bởi vì nó mang tính bối cảnh.

Thử nghiệm phân biệt đối xử: một ván cược bất khả?

Để một sản phẩm được đưa ra thị trường, các nhà cung cấp phải tuân thủ một số thông số kỹ thuật được xác định bên ngoài văn kiện luật, thường dưới dạng các chuẩn mực đã được hài hòa hoá.

Đã có nhiều nước đặt nhiều hy vọng vào các chuẩn mực này, vốn có thể giúp thiết lập các yêu cầu thống nhất về mặt pháp lý và kỹ thuật đối với các hệ thống AI. Ví dụ, các chuẩn mực này có thể xác định các tiêu chí về chất lượng, công bằng, bảo mật hoặc thậm chí còn có khả năng giải thích các hệ thống này, và ngay cả củng cố vị trí chiến lược của châu Âu trong cuộc đua toàn cầu về AI.

Nhưng làm thế nào để phát triển các tiêu chí kỹ thuật trung lập xoay quanh các đánh giá về giá trị đạo đức và văn hóa?

Một số hệ thống trong số này tác động trực tiếp đến quyền không phân biệt đối xử. Ví dụ, việc cảnh sát sử dụng kỹ thuật nhận dạng khuôn mặt đã dẫn đến việc bắt giữ nhiều người da đen, do nhầm lẫn của một hệ thống camera giám sát nhận dạng tự động. Tương tự, thuật toán tuyển dụng của Amazon có nhiều khả năng từ chối một hồ sơ xin việc của phụ nữ một cách dễ dàng hơn, so với một hồ sơ xin việc của nam giới.

Để kiểm định sự hiện diện các dạng thành kiến phân biệt đối xử này, các kỹ thuật hiện tại bao gồm việc kiểm tra mức độ hoạt động thích hợp của hệ thống đối với nhiều phân nhóm dân số khác nhau, tách biệt các cá thể theo giới tính hoặc màu da. Nhưng việc chuẩn mực hóa các phương pháp kiểm định này sẽ làm nảy sinh nhiều khó khăn.

Trước hết, luật pháp của một số nước nghiêm cấm việc xử lý dữ liệu về sắc tộc. Kế đến, việc lựa chọn nhóm người nào để thử nghiệm là một lựa chọn mang tính chính trị: ai sẽ quyết định cần bao nhiêu "màu da" để thử nghiệm? Cuối cùng, một hệ thống như thế sẽ không bao giờ có thể mang tính hoàn toàn công bằng, bởi vì có rất nhiều cách tiếp cận không phân biệt đối xử, mà một số trong số các cách tiếp cận đó không tương thích với nhau.

Mức độ rủi ro phân biệt đối xử "có thể chấp nhận được"?

Do không thể đảm bảo không có phân biệt đối xử dưới mọi góc độ, nên một số lựa chọn sẽ là cần thiết và sẽ phải xác định các ngưỡng dung sai. Khi đó vấn đề đặt ra là: mức độ sai lầm có thể chấp nhận được là gì, và loại sai lầm nào được đề cập đến? Điều này dẫn chúng ta trở lại với câu hỏi: mức độ "chấp nhận được" của rủi ro phân biệt đối xử là bao nhiêu?

Trong các lĩnh vực khác, người ta thường xác định mức độ rủi ro có thể chấp nhận được bằng một tiếp cận định lượng. Ví dụ, đối với mức độ an toàn của một nhà máy điện hạt nhân, rủi ro xảy ra tai nạn sẽ được định lượng, và nhà máy điện có thể mở cửa hoạt động trở lại nếu rủi ro được ghi nhận ở một mức nào đó dưới ngưỡng chấp nhận được. Ngưỡng này không thể bằng 0, bởi vì cách tiếp cận "rủi ro bằng không" sẽ dẫn đến việc loại bỏ hoạt động của nhà máy điện hạt nhân, vốn vả lại cũng mang lại lợi ích cho xã hội. Lợi ích của năng lượng hạt nhân và rủi ro sẽ được cân nhắc, và trên thực tế, ngưỡng rủi ro "có thể chấp nhận được" sẽ được quyết định. Dù ngưỡng này có thể được tranh luận, nhưng nó được hỗ trợ bởi các dữ kiện khoa học: các rủi ro tương đối dễ xác định và đo lường.

Đối với hệ thống AI thì sao? Ví dụ: nếu đề cập trở lại đến vấn đề phân biệt đối xử – phân biệt đối xử về chủng tộc hoặc giới tính, thì một chuẩn mực kỹ thuật không bao giờ có thể cho biết tỷ lệ sai số "tốt" là bao nhiêu, hoặc mức chênh lệch hiệu suất "có thể chấp nhận được" đối với nhiều nhóm dân số khác nhau, bởi vì đáp án phụ thuộc quá nhiều vào bối cảnh và nhận định của con người.

Mục tiêu của các chuẩn mực kỹ thuật trong khuôn khổ AI nên là việc xác định một từ vựng chung, các kiểm định thích hợp, các nghiên cứu về tác động, và một cách chung hơn, các thông lệ thực hành tốt trong suốt vòng đời của các hệ thống AI. Điều này sẽ cung cấp cơ sở để so sánh các hệ thống và thúc đẩy cuộc thảo luận giữa các bên liên quan. Các dạng chuẩn mực này được gọi là "chuẩn mực thông tin", đối lập với các "chuẩn mực chất lượng" hoặc các chuẩn mực về "hiệu suất".

Ví dụ, đối với hệ thống nhận dạng khuôn mặt để đi qua cửa khẩu, một chuẩn mực kỹ thuật có thể mô tả cách thức đo lường độ chính xác của hệ thống và cách thức đo lường sự khác biệt về hiệu suất đối với nhiều nhóm dân số khác nhau. Chuẩn mực này sẽ không xác định mức độ sai số nào và mức độ phân biệt đối xử nào là có thể chấp nhận được, hoặc nhóm dân số nào cần được bảo vệ, bởi vì các lựa chọn này thuộc thẩm quyền của pháp luật và đánh giá của tòa án.

Ai nên quyết định các chuẩn mực?

Các cơ quan chuẩn mực hóa hiện đang làm việc về các chuẩn mực này đối với AI, đặc biệt chú ý đến vấn đề đạo đức. Chúng ta cũng đặc biệt lưu ý đến các sáng kiến của hiệp hội IEEE (Institute of Electrical and Electronics Engineers [Hội Kỹ sư Điện và Điện tử]), những người đầu tiên công bố tiêu chuẩn "xem xét các mối quan tâm về đạo đức khi thiết kế hệ thống", cũng như nhiều tiêu chuẩn khác về tính minh bạch, hoặc thậm chí tính bảo mật dữ liệu.

ISO (International Organization for Standardization [Tổ chức Tiêu chuẩn hóa Quốc tế]) cũng đang soạn thảo các chuẩn mực ISO của họ, nhờ vào nhiều nhóm công tác khác nhau được tổ chức xoay quanh chủ đề chung về AI này, và tập hợp các tổ chức từ nhiều nước khác nhau, chẳng hạn như AFNOR (Association Française de Normalisation [Hiệp hội Tiêu chuẩn hóa của Pháp]) ở Pháp. Công việc này được điều phối ở cấp độ châu Âu bởi CEN-CENELEC (Comité Européen de Normalization [CEN, Ủy ban Tiêu chuẩn hóa châu Âu], Comité Européen de Normalisation Électrotechnique [CENELEC, Ủy ban tiêu chuẩn hóa châu Âu trong lĩnh vực kỹ thuật điện]) và "lộ trình về AI". Tại Hoa Kỳ, Viện Các chuẩn mực và Công nghệ Quốc gia (NIST) chịu trách nhiệm so sánh hiệu suất và tính công bằng của các hệ thống nhận dạng khuôn mặt và, gần đây, đã công bố các hướng dẫn để tránh xảy ra các trường hợp thiên kiến trong các hệ thống tập sự.

Tuy nhiên, những cân nhắc về các vấn đề cơ bản này rất thường diễn ra đằng sau những cánh cửa đóng. Chỉ có những tổ chức được ủy quyền mới có thể phát triển các chuẩn mực này, và dù thường xuyên kêu gọi sự tham gia từ các tổ chức khác bên ngoài, thì khả năng tiếp cận công việc cũng bị hạn chế. Hệ thống hiện tại buộc các công ty phải trả tiền để có được quyền truy cập nội dung các văn bản về chuẩn mực, làm giảm tính minh bạch của quy trình. Khi đó, làm thế nào để có thể đảm bảo các thông số kỹ thuật thực sự bảo vệ các quyền của chúng ta? Liệu có nên cấm việc xây dựng các chuẩn mực không rõ ràng này để ủng hộ một thủ tục cởi mở hơn cho cuộc tranh luận của người dân?

Nhãn hiệu chứng nhận, con dao hai lưỡi

Các chuẩn mực này cũng có thể trở nên vô dụng, thậm chí nguy hiểm, nếu bị sử dụng sai mục đích, chẳng hạn như nếu được sử dụng để dán nhãn là tránh được mọi sự vi phạm quyền tự do của chúng ta. Thật vậy, việc có được nhãn hiệu CE chỉ cho thấy sự tuân thủ các quy tắc của thời điểm hiện tại, do nhà sản xuất tuyên bố, chứ không đảm bảo sự an toàn hoặc không có sự phân biệt đối xử chính thức nào cả.

Vấn đề này càng trầm trọng vì "suy đoán tuân thủ" được tạo ra bởi việc được chứng nhận: khi một hệ thống tuân thủ một chuẩn mực, thì tính an toàn thường ít bị đặt vấn đề. Khi đó, điều cần thiết là phải xây dựng một nhãn hiệu CE không làm giảm trách nhiệm của các nhà cung cấp và của người sử dụng các hệ thống AI. Do đó, việc dán nhãn CE sẽ không đảm bảo một hệ thống sẽ được miễn trừ vi phạm các quyền cơ bản, mà chỉ đảm bảo công ty đã áp dụng các biện pháp để hạn chế các hành vi vi phạm.

Việc dán nhãn CE có thể là một công cụ điều tiết và quản lý tuyệt vời khi kết hợp với các chuẩn mực kỹ thuật, giúp chúng ta nói cùng một ngôn ngữ và so sánh các hệ thống với nhau. Ngược lại, việc quyết định một rủi ro nào là "có thể chấp nhận được" đối với các quyền của con người sẽ không bao giờ có thể được xem là một chuẩn mực kỹ thuật. Cần phải tái khẳng định trách nhiệm của các nhà cung cấp và của người sử dụng các hệ thống này, mà bản thân họ, trong bối cảnh của mình, phải xác định lựa chọn thích hợp nhất để bảo vệ các quyền cơ bản. Quyết định này phải được chứng minh và có thể bị phản bác, nhưng việc phân xử cuối cùng về khả năng chấp nhận mức độ rủi ro phải được giao cho các nhà lập pháp, các cơ quan điều tiết và các thẩm phán.

Tác giả

Winston Maxwell

Mélanie Gornet

[Mélanie Gornet](), nghiên cứu sinh về luật và đạo đức AI, Télécom Paris - Institut Mines-Télécom

Winston Maxwell, Giám đốc Nghiên cứu, Luật và Kỹ thuật số, Télécom Paris - Institut Mines-Télécom

Tuyên bố công khai

Mélanie Gornet có nhận tài trợ từ ANR cho dự án LIMPID (https://anr.fr/Project-ANR-20-CE23-0028) trong khuôn khổ làm luận án tiến sĩ của mình.

Winston Maxwell có nhận tài trợ từ Cơ quan Nghiên cứu Quốc gia

http://www.phantichkinhte123.com/2022/12