NGHIÊN CỨU, PHÁT TRIỂN CÁC HỆ THỐNG TRÍ TUỆ NHÂN TẠO CÓ TRÁCH NHIỆM

Nhằm cung cấp tài liệu và hướng dẫn một số nguyên tắc về nghiên cứu, phát triển các hệ thống trí tuệ nhân tạo (AI) có trách nhiệm, hướng đến một xã hội lấy con người làm trung tâm, mọi người được hưởng những lợi ích từ các hệ thống AI, bảo đảm sự cân bằng hợp lý giữa lợi ích và rủi ro của các hệ thống AI…, ngày 11/06/2024, Bộ Khoa học và Công nghệ (KH&CN) đã ban hành Quyết định số 1290/QĐ-BKHCN hướng dẫn một số nguyên tắc về nghiên cứu, phát triển các hệ thống AI có trách nhiệm. Tạp chí trân trọng giới thiệu 9 nguyên tắc chính được nêu tại Quyết định này.

Ngày 26/01/2021, Thủ tướng Chính phủ đã ký Quyết định số 127/QĐ-TTg ban hành Chiến lược quốc gia về nghiên cứu, phát triển và ứng dụng AI đến năm 2030. Mục tiêu của Chiến lược này là đưa Việt Nam trở thành trung tâm đổi mới sáng tạo và phát triển các giải pháp, ứng dụng AI trong khu vực Đông Nam Á và trên thế giới. Để đạt được mục tiêu này, Thủ tướng Chính phủ đã đề ra 5 nhóm định hướng chiến lược gồm: Xây dựng hệ thống văn bản quy phạm pháp luật liên quan đến AI; xây dựng hạ tầng dữ liệu, tính toán để hỗ trợ nghiên cứu, phát triển và ứng dụng AI; phát triển hệ sinh thái AI; thúc đẩy ứng dụng AI; thúc đẩy hợp tác quốc tế trong lĩnh vực AI. Với những nỗ lực này, Việt Nam hy vọng sẽ xây dựng được 10 thương hiệu AI uy tín trong khu vực và trở thành một điểm sáng về AI trên thế giới.

Sau hơn 2 năm triển khai Chiến lược, Việt Nam đã đạt một số kết quả bước đầu đáng khích lệ. Các sản phẩm, công nghệ AI ngày càng được ứng dụng rộng rãi trong mọi lĩnh vực. Một số tập đoàn, doanh nghiệp của Việt Nam đã quan tâm, đầu tư mạnh mẽ cho AI và từng bước nâng cao khả năng tiếp cận, hấp thụ, làm chủ công nghệ AI.

Cùng với xu thế chung trên thế giới, các hệ thống AI được đánh giá sẽ mang lại nhiều lợi ích to lớn cho con người, xã hội và nền kinh tế Việt Nam thông qua việc hỗ trợ, giải quyết các vấn đề khó khăn mà con người, cộng đồng đang phải đối mặt. Tuy nhiên, việc ứng dụng AI cũng tiềm ẩn nhiều rủi ro, nếu AI bị làm dụng vào mục đích xấu có thể đe dọa nghiêm trọng đến quyền riêng tư và sự an toàn của con người. Nhằm cung cấp tài liệu và hướng dẫn một số nguyên tắc về nghiên cứu, phát triển các hệ thống AI có trách nhiệm, hướng đến một xã hội lấy con người làm trung tâm, mọi người được hưởng những lợi ích từ các hệ thống AI, bảo đảm sự cân bằng hợp lý giữa lợi ích và rủi ro của các hệ thống AI, ngày 11/06/2024, Bộ KH&CN đã ban hành Quyết định số 1290/QĐ-BKHCN về việc hướng dẫn một số nguyên tắc về nghiên cứu, phát triển các hệ thống AI có trách nhiệm. Trong đó nêu rõ 9 nguyên tắc trong nghiên cứu, phát triển các hệ thống AI có trách nhiệm:

Thứ nhất, tinh thần hợp tác, thúc đẩy đổi mới sáng tạo: Các nhà phát triển cần xem xét tính liên kết và khả năng tương tác giữa các hệ thống AI của mình với các hệ thống AI khác thông qua việc xem xét tính đa dạng của các hệ thống AI nhằm: (1) tăng cường lợi ích của hệ thống AI thông qua quá trình kết nối các hệ thống AI; (2) tăng cường sự phối hợp để kiểm soát rủi ro.

Để làm được điều này, các nhà phát triển nên xem xét những điểm sau: i) Tăng cường hợp tác để chia sẻ các thông tin liên quan nhằm bảo đảm tính liên thông, tương tác của hệ thống; ii) Ưu tiên phát triển các hệ thống AI phù hợp các quy chuẩn kỹ thuật, tiêu chuẩn quốc gia hoặc tiêu chuẩn quốc tế (nếu có). Tăng cường chuẩn hóa của các định dạng dữ liệu và tính mở của các giao diện, giao thức trong đó có các giao diện lập trình ứng dụng (API); iii) Quan tâm đến các rủi ro/sự kiện ngoài ý muốn do sự liên kết, tương tác giữa các hệ thống AI; iv) Thúc đẩy việc trao đổi, chia sẻ và minh bạch hóa các thỏa thuận cấp phép, các điều kiện về quyền sở hữu trí tuệ như các bằng sáng chế nhằm góp phần tăng cường tính liên kết và khả năng tương tác giữa các hệ thống AI khi liên quan đến các tài sản trí tuệ (không liên quan đến bí mật kinh doanh); v) Đóng góp vào việc duy trì sự phát triển kinh tế bền vững và giải quyết các thách thức của nền kinh tế, xã hội; vi) Thúc đẩy sự hợp tác trong các ngành, lĩnh vực và các bên có liên quan nhằm phát triển cộng đồng AI ở Việt Nam.

Thứ hai, tính minh bạch: Nhà phát triển cần chú ý đến việc kiểm soát đầu vào/đầu ra của hệ thống AI và khả năng giải thích các phân tích có liên quan. Theo đó, các hệ thống AI tuân theo nguyên tắc này thường là các hệ thống có thể ảnh hưởng đến tính mạng, thân thể, quyền riêng tư hoặc tài sản của người dùng hoặc bên thứ ba liên quan. Khi đó, các nhà phát triển cần chú ý đến khả năng xác định rõ các đầu vào và đầu ra của hệ thống AI cũng như khả năng giải thích liên quan dựa trên các đặc điểm của công nghệ được áp dụng và cách sử dụng chúng để bảo đảm có sự tin tưởng của xã hội, bao gồm cả người dùng.

Thứ ba, khả năng kiểm soát: Để đánh giá các rủi ro liên quan đến khả năng kiểm soát của hệ thống AI, các nhà phát triển cần thực hiện đánh giá trước (là quá trình đánh giá liệu hệ thống có đáp ứng với các yêu cầu kỹ thuật và tiêu chuẩn tương ứng). Một trong những phương pháp đánh giá rủi ro là tiến hành thử nghiệm trong một không gian riêng như trong phòng thí nghiệm hoặc môi trường thử nghiệm nơi đã có các biện pháp bảo đảm an ninh, an toàn trước khi đưa vào áp dụng thực tế. Ngoài ra, để bảo đảm khả năng kiểm soát hệ thống AI, các nhà phát triển nên chú ý đến việc giám sát hệ thống (có công cụ đánh giá/giám sát hoặc hiệu chỉnh/cập nhật dựa trên các phản hồi của người dùng) và các biện pháp ứng phó (như ngắt hệ thống, ngắt mạng…) được thực hiện bởi con người hay các hệ thống AI đáng tin cậy khác.

Thứ tư, an toàn: Nhà phát triển cần bảo đảm rằng hệ thống AI sẽ không gây tổn hại đến tính mạng, thân thể hoặc tài sản của người dùng hoặc bên thứ ba kể cả thông qua trung gian. Về cơ bản, khuyến khích nhà phát triển tham khảo các tiêu chuẩn quốc tế có liên quan và chú ý đến những điểm sau đây, trong đó đặc biệt lưu ý các khả năng đầu ra hoặc chương trình thay đổi do quá trình huấn luyện hệ thống AI: i) Tiến hành đánh giá trước nhằm xác định và giảm thiểu các rủi ro liên quan đến sự an toàn của hệ thống AI; ii) Trong suốt các giai đoạn phát triển của hệ thống AI, thực hiện các biện pháp nhằm bảo đảm an toàn nội tại (giảm các yếu tố rủi ro như mức năng lượng của các thiết bị tạo ra sự kiện…) và an toàn chức năng (giảm thiểu rủi ro bằng cách sử dụng các thiết bị điều khiển bổ sung như tự động dừng khi có sự cố…); iii) Giải thích ý tưởng/ý định của người thiết kế hệ thống AI và sự phù hợp cho các bên liên quan; việc thực hiện

đánh giá sự an toàn đối với tính mạng, thân thể hoặc tài sản của người dùng và bên thứ ba (ví dụ như những ý tưởng để ưu tiên bảo vệ tính mạng, thân thể, tài sản của con người khi xảy ra tai nạn với robot được trang bị AI).

Thứ năm, bảo mật: Bên cạnh việc tuân thủ các văn bản, hướng dẫn và thực hiện các biện pháp bảo mật thông tin theo quy định (của các cơ quan chuyên môn, có thẩm quyền), các nhà phát triển cần chú ý đến những điểm sau đây: i) Cần chú ý đến độ tin cậy (nghĩa là liệu các hoạt động có được thực hiện như dự định và không bị ảnh hưởng bởi bên thứ ba một cách bất hợp pháp) và khả năng chống chịu các dạng tấn công hoặc tai nạn vật lý của hệ thống AI; và đồng thời cần bảo đảm tính bảo mật, sự toàn vẹn và tính khả dụng của các thông tin cần thiết liên quan đến sự an toàn thông tin của hệ thống AI; ii) Thực hiện đánh giá trước trước nhằm xác định và kiểm soát các rủi ro liên quan đến an toàn của hệ thống AI; iii) Thực hiện các biện pháp cần thiết để duy trì tính bảo mật trong phạm vi có thể dựa trên đặc điểm của các công nghệ được áp dụng trong suốt quá trình phát triển hệ thống AI (bảo mật theo thiết kế).

Thứ sáu, quyền riêng tư: Nhà phát triển cần bảo đảm rằng hệ thống AI không vi phạm quyền riêng tư của người dùng hoặc bên thứ ba. Quyền riêng tư được đề cập trong nguyên tắc này bao gồm quyền riêng tư về không gian (sự yên bình trong cuộc sống cá nhân), quyền riêng tư về thông tin (dữ liệu cá nhân) và sự bí mật của việc thông tin liên lạc. Các nhà phát triển cần áp dụng các quy định, hướng dẫn hiện hành (của cơ quan chức năng, cơ quan có thẩm quyền); có thể tham khảo các tiêu chuẩn, hướng dẫn quốc tế về quyền riêng tư; và thực hiện các thêm hướng dẫn sau đây, trong đó đặc biệt lưu ý các khả năng đầu ra hoặc chương trình thay đổi do quá trình huấn luyện hệ thống AI: i) Thực hiện đánh giá trước các rủi ro xâm phạm quyền riêng tư và tiến hành đánh giá trước các tác động đến quyền riêng tư (từ khi thiết kế); ii) Trong phạm vi có thể, thực hiện các biện pháp phù hợp với đặc điểm của công nghệ được áp dụng trong suốt quá trình phát triển hệ thống AI (từ khi thiết kế) để tránh xâm phạm quyền riêng tư khi đưa vào sử dụng.

Thứ bảy, tôn trọng quyền và phẩm giá con người: Khi phát triển các hệ thống AI có liên quan tới con người, các nhà phát triển phải đặc biệt quan tâm đến việc tôn trọng quyền và phẩm giá con người của các cá nhân liên quan. Trong phạm vi có thể, tùy theo đặc điểm của công nghệ được áp dụng, các nhà phát triển cần thực hiện các biện pháp để bảo đảm không gây ra sự phân biệt đối xử, không công bằng do thiên vị (định kiến) trong dữ liệu khi huấn luyện hệ thống AI. Các nhà phát triển cần thực hiện các biện pháp phòng ngừa để bảo đảm rằng hệ thống AI không vi phạm các giá trị của con người, đạo đức xã hội theo các nguyên tắc cơ bản của Việt Nam (ví dụ, các giá trị bao gồm yêu nước, đoàn kết, tự cường, nghĩa tình, trung thực, trách nhiệm, kỷ cương, sáng tạo…).

Thứ tám, hỗ trợ người dùng: Nhà phát triển cần bảo đảm rằng hệ thống AI sẽ hỗ trợ người dùng và tạo điều kiện cho họ cơ hội lựa chọn theo cách phù hợp. Để hỗ trợ người dùng, các nhà phát triển hệ thống AI cần chú ý các điểm sau đây: i) Tạo ra các giao diện sẵn sàng để cung cấp thông tin kịp thời và phù hợp nhằm giúp người dùng đưa ra quyết định và sử dụng thuận tiện; ii) Xem xét cung cấp các chức năng cho người dùng cơ hội lựa chọn kịp thời và phù hợp (ví dụ, các cài đặt mặc định, các tùy chọn dễ hiểu, phản hồi, cảnh báo khẩn cấp, xử lý lỗi); iii) Thực hiện các biện pháp giúp hệ thống AI dễ sử dụng hơn cho những người dễ bị tổn thương trong xã hội (người già, người khuyết tật). Ngoài ra, các nhà phát triển nên cung cấp cho người dùng các thông tin cần thiết trong đó lưu ý các khả năng đầu ra hoặc chương trình thay đổi do quá trình huấn luyện hệ thống AI; và hướng dẫn người sử dụng cách thức sử dụng hệ thống AI rõ ràng để tránh xảy ra nguy hiểm không mong muốn (như các điều kiện sử dụng hay các biện pháp giảm thiểu rủi ro…).

Thứ chín, trách nhiệm giải trình: Các nhà phát triển cần thực hiện trách nhiệm giải trình đối với các hệ thống AI mà họ đã phát triển để bảo đảm niềm tin của người dùng. Cụ thể, các nhà phát triển cần cung cấp cho người dùng thông tin để giúp họ lựa chọn và sử dụng hệ thống AI. Ngoài ra, để tăng sự chấp nhận của xã hội đối với các hệ thống AI, bao gồm cả người dùng, sau khi thực hiện các hướng dẫn nêu trên, các nhà phát triển nên thực hiện thêm: i) cung cấp cho người dùng thông tin và mô tả về đặc tính kỹ thuật của hệ thống AI mà họ phát triển, các thuật toán, các cơ chế bảo đảm an toàn… và ii) lắng nghe các quan điểm và đối thoại với các bên liên quan. Ngoài ra, các nhà phát triển cũng cần thực hiện chia sẻ thông tin và hợp tác chặt chẽ với các nhà cung cấp để bảo đảm cập nhật và giải quyết các vấn đề liên quan trong quá trình cung cấp dịch vụ và sử dụng các hệ thống AI

PT *(tổng hợp)*