BỘ KHOA HỌC VÀ CÔNG NGHỆ

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

HƯỚNG DẪN
MỘT SỐ NGUYÊN TẮC VỀ NGHIÊN CỨU, PHÁT TRIỂN CÁC HỆ THỐNG TRÍ TUỆ NHÂN TẠO CÓ TRÁCH NHIỆM (PHIÊN BẢN 1.0)

(kèm theo Quyết định số 1290 /QĐ-BKHCN ngày 11 tháng 6 năm 2024 của Bộ trưởng Bộ Khoa học và Công nghệ)

1. Tổng quan

Theo xu thế chung trên thế giới, các hệ thống trí tuệ nhân tạo (TTNT) được đánh giá sẽ mang lại các 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, song song với quá trình đó, cần nghiên cứu, có biện pháp giảm thiểu các rủi ro trong quá trình phát triển, sử dụng TTNT; và cân đối các yếu tố kinh tế, đạo đức và pháp lý liên quan. Vì vậy, các cơ quan chuyên môn cần nghiên cứu, xây dựng các tiêu chuẩn, hướng dẫn để định hướng kể cả đó là các quy định mềm và không có tính ràng buộc. Bên cạnh đó, việc chia sẻ, trao đổi thông tin về các quy trình, các biện pháp thực hành tốt giữa các bên liên quan (như nhà phát triển, nhà cung cấp dịch vụ, người dùng) cũng sẽ thúc đẩy sự đồng thuận để gia tăng lợi ích từ các hệ thống TTNT và kiểm soát được các rủi ro.

Trên tinh thần đó, việc nghiên cứu, phát triển các hệ thống TTNT ở Việt Nam cần dựa trên các quan điểm cơ bản như sau:

- *Thứ nhất,* hướng đến một xã hội lấy con người làm trung tâm, mọi người được hưởng những lợi ích từ cuộc sống cũng như từ các hệ thống TTNT.

- *Thứ hai,* đảm bảo sự cân bằng hợp lý giữa lợi ích và rủi ro của các hệ thống TTNT, cụ thể là: (1) phát huy lợi ích của TTNT thông qua các hoạt động nghiên cứu, phát triển và đổi mới sáng tạo; và (2) giảm thiểu nguy cơ xâm phạm quyền hoặc lợi ích hợp pháp của các tổ chức, cá nhân từ các hệ thống TTNT.

- *Thứ ba,* đảm bảo các hoạt động nghiên cứu, phát triển các hệ thống TTNT dựa trên các công nghệ hoặc kỹ thuật cụ thể nhưng vẫn đảm bảo tính trung lập về công nghệ và các nhà phát triển cũng không bị ảnh hưởng bởi sự phát triển quá nhanh của các công nghệ liên quan đến TTNT trong tương lai.

- *Thứ tư,* ở giai đoạn hiện nay, tạm thời xác định rằng các văn bản có thể ở dạng hướng dẫn, không có tính ràng buộc và khuyến khích xây dựng, áp dụng các tiêu chuẩn, quy trình thực hành dựa trên các khuyến nghị quốc tế làm nền tảng để thúc đẩy nghiên cứu, phát triển và sử dụng các hệ thống TTNT.

- *Thứ năm,* trong mọi trường hợp, khuyến khích việc trao đổi, thảo luận với sự tham gia của các bên liên quan đến hệ thống TTNT cho dù việc nghiên cứu, phát triển các hệ thống TTNT trong các lĩnh vực có các đặc điểm, cách thức sử dụng và

lợi ích, rủi ro khác nhau.

- *Thứ sáu,* các nguyên tắc, hướng dẫn sẽ tiếp tục được nghiên cứu, cập nhật để phù hợp với tình hình thực tiễn.

2. Mục tiêu

Tài liệu hướng dẫn này nhằm:

- Thúc đẩy sự quan tâm của các bên liên quan trong việc nghiên cứu, phát triển và sử dụng các hệ thống/ứng dụng TTNT ở Việt Nam một cách có trách nhiệm.

- Thúc đẩy việc nghiên cứu, phát triển và sử dụng các hệ thống/ứng dụng TTNT một cách an toàn và có trách nhiệm, đồng thời hạn chế tối đa các ảnh hưởng tiêu cực cho con người và cộng đồng.

- Thúc đẩy việc chia sẻ kinh nghiệm trong hoạt động nghiên cứu, phát triển và sử dụng các hệ thống/ứng dụng TTNT nhằm đạt được sự tin tưởng của người dùng và xã hội đối với TTNT cũng chính là tạo điều kiện thuận lợi cho việc nghiên cứu, phát triển TTNT ở Việt Nam.

3. Phạm vi

Tài liệu hướng dẫn này nêu ra một số nguyên tắc chung cần chú ý trong nghiên cứu, phát triển các hệ thống TTNT một cách có trách nhiệm và khuyến nghị tự nguyện tham khảo, áp dụng trong quá trình nghiên cứu, thiết kế, phát triển, cung cấp các hệ thống TTNT.

4. Đối tượng

Cơ quan, tổ chức khoa học và công nghệ, tổ chức, doanh nghiệp, cá nhân có hoạt động nghiên cứu, thiết kế, phát triển, cung cấp các hệ thống TTNT được khuyến khích áp dụng các nội dung trong tài liệu hướng dẫn này.

5. Khái niệm, thuật ngữ

5.1. Trí tuệ nhân tạo (Artificial intelligence – AI) là công nghệ hướng đến việc mô phỏng trí thông minh của con người bằng cách sử dụng máy móc, đặc biệt là các hệ thống máy tính.

5.2. Hệ thống TTNT (AI system) là hệ thống kỹ thuật tạo ra các kết quả đầu ra như nội dung, dự báo, khuyến nghị hoặc quyết định cho một tập hợp các mục tiêu xác định bởi con người (Đối với hệ thống kỹ thuật, các mô hình, biểu diễn dữ liệu, tri thức, quy trình... được sử dụng để tiến hành các tác vụ có thể được phát triển bằng nhiều kỹ thuật và phương pháp tiếp cận khác nhau liên quan đến trí tuệ nhân tạo; Hệ thống TTNT có thể được thiết kế để hoạt động ở mức độ tự động hóa khác nhau).

5.3. Mô hình (Model) là dạng biểu diễn vật lý, toán học hoặc logic khác của một hệ thống, thực thể, hiện tượng, quá trình, dữ liệu.

5.4. Trách nhiệm giải trình (Accountability) là trạng thái có thể trả lời cho các hành động, quyết định và hiệu năng (của hệ thống TTNT).

5.5. Tính minh bạch (Transparency) của một hệ thống là thuộc tính của một

hệ thống mà thông tin phù hợp về hệ thống tạo ra để khả dụng đối với các bên liên quan có liên đới đến thông tin đó (Thông tin phù hợp cho tính minh bạch của hệ thống có thể bao gồm các khía cạnh như tính năng, hiệu năng, các giới hạn, thành phần, thủ tục, biện pháp, các mục tiêu thiết kế, lựa chọn thiết kế và giả định, các nguồn dữ liệu và giao thức gắn nhãn; Tiết lộ không phù hợp về một số khía cạnh của hệ thống có thể vi phạm các yêu cầu về bảo mật, quyền riêng tư hoặc các yêu cầu về tính bảo mật).

5.6. Rủi ro (Risk) là sự ảnh hưởng của tính bất định đến các mục tiêu.

5.7. Thiên vị (Bias) là sự khác biệt mang tính hệ thống trong cách đối xử với một số đối tượng, người hoặc nhóm nhất định so với các đối tượng, người hoặc nhóm khác (cách đối xử bao gồm nhận thức, quan sát, dự đoán hoặc quyết định).

5.8. Nhà phát triển/Đơn vị/Tổ chức phát triển (Developer) là những người thực hiện hoạt động nghiên cứu, phát triển, cung cấp các hệ thống TTNT (*phần tiếp theo của tài liệu hướng dẫn này gọi chung là "Nhà phát triển"*).

5.9. Người dùng/Cá nhân sử dụng (User) là những người sử dụng hệ thống TTNT, bao gồm người dùng cuối hoặc nhà cung cấp thuộc bên thứ ba thực hiện cung cấp các dịch vụ/hệ thống TTNT (*phần tiếp theo của tài liệu hướng dẫn này gọi chung là "Người dùng"*).

5.10. Bên liên quan (Stakeholder) là bất kỳ cá nhân, nhóm hoặc tổ chức nào có thể ảnh hưởng, bị ảnh hưởng hoặc tự nhận thức bị ảnh hưởng bởi một quyết định hoặc hành động.

6. Các nguyên tắc nghiên cứu, phát triển các hệ thống TTNT có trách nhiệm và hướng dẫn thực hiện

6.1. Tinh thần hợp tác, thúc đẩy đổi mới sáng tạo

Nhà phát triển cần chú ý đến khả năng kết nối và tương tác của các hệ thống TTNT. Cụ thể, các nhà phát triển cần xem xét tính liên kết và khả năng tương tác giữa các hệ thống TTNT của mình với các hệ thống TTNT khác thông qua việc xem xét tính đa dạng của các hệ thống TTNT nhằm: (1) tăng cường lợi ích của hệ thống TTNT thông qua quá trình kết nối các hệ thống TTNT; và (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:

- Tăng cường hợp tác để chia sẻ các thông tin liên quan nhằm đảm bảo tính liên thông, tương tác của hệ thống.

- Ưu tiên phát triển các hệ thống TTNT 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).

- Quan tâm đến các rủi ro/sự kiện ngoài ý muốn do sự liên kết, tương tác giữa các hệ thống TTNT.

- 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 TTNT khi liên quan đến các tài sản trí tuệ (không liên quan đến bí mật kinh doanh).

- 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.

- Thúc đẩy sự hợp tác trong các ngành, lĩnh vực và các bên có liên quan nhằm phát triển cộng đồng TTNT ở Việt Nam.

6.2. Tính minh bạch

Nhà phát triển cần chú ý đến việc kiểm soát đầu vào/đầu ra của hệ thống TTNT và khả năng giải thích các phân tích có liên quan. Theo đó, các hệ thống TTNT 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 TTNT 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 đảm bảo có sự tin tưởng của xã hội, bao gồm cả người dùng.

6.3. Khả năng kiểm soát

Nhà phát triển cần chú ý đến khả năng kiểm soát hệ thống TTNT. Để đánh giá các rủi ro liên quan đến khả năng kiểm soát của hệ thống TTNT, 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 đảm bảo an ninh, an toàn trước khi đưa vào áp dụng thực tế.

Ngoài ra, để đảm bảo khả năng kiểm soát hệ thống TTNT, 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 TTNT đáng tin cậy khác.

6.4. An toàn

Nhà phát triển cần đảm bảo rằng hệ thống TTNT 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 TTNT:

- Tiến hành đánh giá trước nhằm xác định và giảm thiểu các rủi ro liên quan đến sự an toàn của hệ thống TTNT.

- Trong suốt các giai đoạn phát triển của hệ thống TTNT, thực hiện các biện pháp nhằm đảm bảo 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...).

- Giải thích ý tưởng/ý định của người thiết kế hệ thống TTNT 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 TTNT).

6.5. Bảo mật

Các nhà phát triển cần chú ý đến tính bảo mật của hệ thống TTNT. 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, trong đó đặc biệt lưu ý các khả năng đầu ra hoặc chương trình thay đổi do quá trình huấn luyện hệ thống TTNT:

- 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 TTNT; và đồng thời cần đảm bảo: (1) tính bảo mật; (2) sự toàn vẹn; và (3) tính khả dụng của các thông tin cần thiết liên quan đến sự an toàn thông tin của hệ thống TTNT.

- Thực hiện đánh giá trước trước nhằm xác định và kiểm soát các rủi ro liên quan đến an toàn của hệ thống TTNT.

- 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 TTNT (bảo mật theo thiết kế).

6.6. Quyền riêng tư

Nhà phát triển cần đảm bảo rằng hệ thống TTNT 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 TTNT:

- 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).

- 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 TTNT (từ khi thiết kế) để tránh xâm phạm quyền riêng tư khi đưa vào sử dụng.

6.7. Tôn trọng quyền và phẩm giá con người

Khi phát triển các hệ thống TTNT 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 đảm bảo không gây ra

sự phân biệt đối xử, không công bằng do thiên vị (định kiến) trong dữ liệu khi huấn luyện hệ thống TTNT.

Các nhà phát triển cần thực hiện các biện pháp phòng ngừa để đảm bảo rằng hệ thống TTNT 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…).

6.8. Hỗ trợ người dùng

Nhà phát triển cần đảm bảo rằng hệ thống TTNT 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 TTNT cần chú ý các điểm sau đây:

- 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.

- 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).

- Thực hiện các biện pháp giúp hệ thống TTNT 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 TTNT; và hướng dẫn người sử dụng cách thức sử dụng hệ thống TTNT 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…).

6.9. Trách nhiệm giải trình

Nhà phát triển cần thực hiện trách nhiệm giải trình của mình đối với các bên liên quan bao gồm cả người dùng hệ thống TTNT. Các nhà phát triển cần thực hiện trách nhiệm giải trình đối với các hệ thống TTNT mà họ đã phát triển để đảm bảo 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 TTNT. Ngoài ra, để tăng sự chấp nhận của xã hội đối với các hệ thống TTNT, bao gồm cả người dùng, sau khi thực hiện các hướng dẫn nêu trên, các nhà phát triển nên thực hiện thêm: (1) cung cấp cho người dùng thông tin và mô tả về đặc tính kỹ thuật của hệ thống TTNT mà họ phát triển, các thuật toán, các cơ chế đảm bảo an toàn…; và (2) 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 đảm bảo cập nhật và giải quyết các vấn đề liên quan trong quá trình cung cấp dịch vụ và sử dụng các hệ thống TTNT./.