# Đề xuất hướng dẫn Khung đạo đức trí tuệ nhân tạo quốc gia

**Bộ Khoa học và Công nghệ đang lấy ý kiến đối với dự thảo Thông tư hướng dẫn Khung đạo đức trí tuệ nhân tạo quốc gia.**




Ảnh minh họa (Nguồn internet)

Dự thảo Thông tư này áp dụng đối với cơ quan quản lý nhà nước; tổ chức, cá nhân tham gia thiết kế, phát triển, cung cấp, triển khai, vận hành hoặc sử dụng hệ thống trí tuệ nhân tạo phục vụ hoạt động quản lý nhà nước và cung cấp dịch vụ công.

Thông tư này là căn cứ để các tổ chức, doanh nghiệp tham chiếu xây dựng quy tắc ứng xử nội bộ và thực hiện tự đánh giá mức độ tuân thủ các nguyên tắc đạo đức trí tuệ nhân tạo trong các hoạt động có liên quan.

## *Các tiêu chí trong Khung đạo đức trí tuệ nhân tạo quốc gia*

Khung đạo đức trí tuệ nhân tạo quốc gia được phát triển bảo đảm các tiêu chí sau:

1. Để bảo đảm việc sử dụng hệ thống trí tuệ nhân tạo an toàn, tin cậy và không gây hại đến tính mạng, sức khỏe, danh dự, nhân phẩm và đời sống tinh thần của con người, hệ thống trí tuệ nhân tạo phải tuân thủ các tiêu chí sau:

An toàn kỹ thuật và an ninh mạng: Hệ thống phải có khả năng chống chịu rủi ro, bảo vệ tính toàn vẹn của dữ liệu huấn luyện, chống lại các hành vi đầu độc dữ liệu (Data Poisoning), ngăn ngừa sự cố và các cuộc tấn công mạng nhằm bảo vệ tính mạng, sức khỏe và tài sản của người dùng.

Sự kiểm soát của con người: Hệ thống phải duy trì sự kiểm soát và khả năng can thiệp của con người; con người có thể giám sát, tạm dừng hoặc vô hiệu hóa hệ thống khi có nguy cơ gây hại.

Bảo vệ danh dự, nhân phẩm: Hệ thống không được tạo ra hoặc lan truyền nội dung xúc phạm danh dự, nhân phẩm, hoặc gây tổn hại đến đời sống tinh thần của con người.

2. Để đảm bảo kết quả xử lý của hệ thống trí tuệ nhân tạo tôn trọng quyền con người, quyền công dân, bảo đảm công bằng, minh bạch và không phân biệt đối xử; hệ thống trí tuệ nhân tạo phải tuân thủ các tiêu chí sau:

Tôn trọng quyền con người và quyền riêng tư: Bảo đảm tuân thủ Hiến pháp và pháp luật về bảo vệ dữ liệu cá nhân; không xâm phạm trái phép đời tư của công dân.

Công bằng và không phân biệt đối xử: Dữ liệu và thuật toán phải được rà soát để loại bỏ thiên kiến (bias); không gây phân biệt đối xử dựa trên giới tính, dân tộc, tôn giáo, địa vị xã hội hoặc các đặc điểm khác được pháp luật bảo vệ.

Minh bạch và giải trình: Việc sử dụng trí tuệ nhân tạo phải được công khai, minh bạch, bảo đảm người dùng được thông báo và có thể nhận biết rõ khi nội dung, dịch vụ hoặc quyết định có sự tham gia của hệ thống trí tuệ nhân tạo; các chủ thể liên quan phải có khả năng giải trình về quá trình hoạt động, quyết định và các hệ quả phát sinh từ hệ thống trí tuệ nhân tạo.

Đối với nội dung do hệ thống trí tuệ nhân tạo tạo ra hoặc chỉnh sửa, nhà cung cấp và bên triển khai có trách nhiệm thông báo, gắn nhãn hoặc áp dụng biện pháp phù hợp để người tiếp nhận nhận biết, đồng thời thực hiện đầy đủ các nghĩa vụ về minh bạch theo quy định của pháp luật về trí tuệ nhân tạo và pháp luật có liên quan.

3. Để việc sử dụng hệ thống trí tuệ nhân tạo sẽ góp phần thúc đẩy hạnh phúc, thịnh vượng và hướng tới sự phát triển bền vững của con người, cộng đồng và xã hội, hệ thống trí tuệ nhân tạo phải tuân thủ các tiêu chí sau:

Phúc lợi xã hội: Góp phần nâng cao chất lượng cuộc sống, hỗ trợ cộng đồng và thúc đẩy sự thịnh vượng chung.

Phát triển bền vững và AI Xanh: Ưu tiên phát triển các công nghệ sử dụng hiệu quả năng lượng, tiết kiệm tài nguyên và giảm thiểu tác động tiêu cực đến môi trường.

Tính bao trùm: Hỗ trợ thu hẹp khoảng cách số, tạo điều kiện cho các nhóm yếu thế tiếp cận và thụ hưởng lợi ích từ trí tuệ nhân tạo.

4. Để hoạt động nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo khuyến khích sáng tạo, đổi mới và trách nhiệm xã hội trong nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo, hệ thống trí tuệ nhân tạo phải tuân thủ các tiêu chí:

Đổi mới sáng tạo có trách nhiệm: Khuyến khích tìm tòi giải pháp công nghệ mới nhưng phải gắn liền với đánh giá rủi ro và đạo đức ngay từ khâu thiết kế (Ethics by Design).

Tuân thủ trong nghiên cứu và phát triển: Quá trình thu thập, xử lý dữ liệu để huấn luyện, kiểm thử phải hợp pháp; bảo đảm quyền sở hữu trí tuệ và an ninh mạng.

Trách nhiệm trong ứng dụng: Tổ chức, cá nhân ứng dụng AI phải chịu trách nhiệm về kết quả sử dụng; không sử dụng AI để xâm phạm lợi ích quốc gia, công cộng hoặc quyền lợi hợp pháp của người khác.

## *Nguyên tắc đánh giá đạo đức trí tuệ nhân tạo*

Việc đánh giá hệ thống trí tuệ nhân tạo tuân thủ các tiêu chí dựa trên:

Hồ sơ, tài liệu (Document Review): Rà soát tính đầy đủ, hợp lệ của các chính sách, quy trình, tài liệu thiết kế và nhật ký vận hành.

Kiểm thử kỹ thuật (Technical Testing): Sử dụng các công cụ, kịch bản để kiểm tra thực tế hoạt động của hệ thống, bao gồm dữ liệu đầu vào, thuật toán và kết quả đầu ra.

Đánh giá quy trình (Process Audit): Xem xét cách thức tổ chức, vận hành và giám sát hệ thống trong thực tế so với quy định nội bộ.

Việc Bộ Khoa học và Công nghệ ban hành Thông tư Khung đạo đức trí tuệ nhân tạo quốc gia để cập nhật, thay thế Quyết định 1290/ QĐ-BKHCN về việc hướng dẫn một số nguyên tắc về nghiên cứu, phát triển các

hệ thống Trí tuệ nhân tạo có trách nhiệm; tăng tính pháp lý, song hành cùng các quy định về phương pháp đánh giá, các ngưỡng tối thiểu để đảm bảo hệ thống trí tuệ nhân tạo hoặc các sản phẩm có tích hợp hệ thống trí tuệ nhân tạo tuân thủ các tiêu chí trong Khung đạo đức quốc gia là rất cần thiết.

**Trung tâm Truyền thông KH&CN**

*Nguồn: https://mst.gov.vn.*