BỘ KHOA HỌC VÀ CÔNG NGHỆ	CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM Độc lập - Tự do - Hạnh phúc
Số: 05 /2026/TT-BKHCN	*Hà Nội, ngày 10 tháng 3 năm 2026*

THÔNG TƯ
Ban hành Khung đạo đức trí tuệ nhân tạo quốc gia

Căn cứ Luật Trí tuệ nhân tạo số 134/2025/QH15 ngày 10 tháng 12 năm 2025;

Căn cứ Nghị định số 55/2025/NĐ-CP ngày 02 tháng 3 năm 2025 của Chính phủ quy định chức năng, nhiệm vụ, quyền hạn và cơ cấu tổ chức của Bộ Khoa học và Công nghệ;

Theo đề nghị của Viện trưởng Viện Công nghệ số và Chuyển đổi số quốc gia và Vụ trưởng Vụ Pháp chế,

Bộ trưởng Bộ Khoa học và Công nghệ ban hành Thông tư ban hành Khung đạo đức trí tuệ nhân tạo quốc gia.

Điều 1. Phạm vi điều chỉnh và đối tượng áp dụng

1. Thông tư này ban hành Khung đạo đức trí tuệ nhân tạo quốc gia (sau đây gọi tắt là Khung đạo đức).

2. Thông tư này áp dụng đối với cơ quan quản lý nhà nước, tổ chức, cá nhân dưới vai trò nhà phát triển, nhà cung cấp, bên triển khai hoặc người sử dụng hệ thống trí tuệ nhân tạo phục vụ hoạt động quản lý nhà nước hoặc cung cấp dịch vụ công.

3. Tổ chức, cá nhân Việt Nam và nước ngoài có hoạt động nghiên cứu, phát triển, cung cấp, triển khai hoặc sử dụng hệ thống trí tuệ nhân tạo tại Việt Nam được khuyến khích áp dụng Thông tư này.

Điều 2. Giải thích từ ngữ

Trong Thông tư này, các từ ngữ dưới đây được hiểu như sau:

1. *Đạo đức trí tuệ nhân tạo* là hệ giá trị, nguyên tắc và chuẩn mực định hướng việc nghiên cứu, phát triển, cung cấp, triển khai và sử dụng trí tuệ nhân tạo nhằm bảo đảm tôn trọng con người, quyền con người, lợi ích công cộng và phát triển bền vững.

2. *Đầu độc dữ liệu* là các cuộc tấn công nhằm sửa đổi, thao túng tập dữ liệu huấn luyện.

3. *Đầu độc mô hình* là tấn công nhắm vào các thành phần tiền huấn luyện của mô hình được dùng trong quá trình huấn luyện.

Điều 3. Khung đạo đức trí tuệ nhân tạo quốc gia

1. Bảo đảm an toàn, độ tin cậy và không gây hại đến tính mạng, sức khỏe, danh dự, nhân phẩm và đời sống tinh thần của con người.

a) Thiết kế an toàn ngay từ đầu: Tổ chức, cá nhân có trách nhiệm xác định trước các kịch bản gây hại có thể xảy ra đối với tính mạng, sức khỏe, danh dự, nhân phẩm và xây dựng biện pháp phòng ngừa.

b) Độ tin cậy và chất lượng: Tổ chức, cá nhân có trách nhiệm thiết lập tiêu chí chất lượng dữ liệu, mô hình, kết quả đầu ra; thực hiện xây dựng các cơ chế kiểm thử, xác nhận, kiểm định nội bộ trước khi triển khai.

c) Kiểm soát của con người: Tổ chức, cá nhân có trách nhiệm xây dựng cơ chế giám sát và can thiệp của con người phù hợp với mức độ ảnh hưởng của hệ thống; bảo đảm duy trì sự kiểm soát và khả năng can thiệp của con người đối với mọi quyết định và hành vi của hệ thống trí tuệ nhân tạo.

d) Khả năng phục hồi, ứng phó và bảo mật: Tổ chức, cá nhân xây dựng cơ chế tiếp nhận phản ánh, phát hiện lỗi và khắc phục; có kế hoạch dự phòng trong trường hợp hệ thống hoạt động sai lệch hoặc bị lạm dụng.

e) Bảo đảm an ninh của hệ thống trí tuệ nhân tạo: Tổ chức, cá nhân áp dụng biện pháp bảo vệ phù hợp để phòng ngừa, phát hiện, ngăn chặn và ứng phó với các hành vi xâm nhập, chiếm quyền điều khiển, đầu độc dữ liệu, đầu độc mô hình, tấn công đối nghịch, khai thác lỗ hổng, rò rỉ dữ liệu và lạm dụng hệ thống trí tuệ nhân tạo; bảo đảm tính bí mật, toàn vẹn và sẵn sàng của dữ liệu, mô hình, thuật toán và hạ tầng liên quan.

2. Tôn trọng quyền con người, quyền công dân, bảo đảm công bằng, minh bạch và không phân biệt đối xử trong phát triển và sử dụng trí tuệ nhân tạo.

a) Tôn trọng quyền con người, quyền công dân: Tổ chức, cá nhân áp dụng biện pháp rà soát phù hợp để bảo đảm hệ thống trí tuệ nhân tạo không xâm phạm quyền riêng tư, dữ liệu cá nhân, tự do ý chí, quyền tiếp cận thông tin, quyền được đối xử bình đẳng và các quyền hợp pháp khác theo quy định của pháp luật.

b) Công bằng và không phân biệt đối xử: Tổ chức, cá nhân sử dụng các biện pháp nhận diện và giảm thiểu thiên lệch dữ liệu, thiên lệch mô hình và thiên lệch vận hành; bảo đảm xem xét đầy đủ tác động đến nhóm dễ bị tổn thương bao gồm trẻ em, người cao tuổi, người khuyết tật, nhóm yếu thế.

c) Minh bạch: Tổ chức, cá nhân xây dựng thông báo phù hợp về việc có sử dụng trí tuệ nhân tạo; bảo đảm cung cấp thông tin ở mức hợp lý về mục tiêu, phạm

vi, dữ liệu, cách thức hoạt động tổng quát và giới hạn của hệ thống; bảo đảm không gây hiểu nhầm về năng lực của hệ thống.

d) Khả năng giải thích và trách nhiệm giải trình: Tổ chức, cá nhân xác định rõ các tác động mà hệ thống có thể gây ra, chuẩn bị tài liệu giải thích và bằng chứng về quá trình thiết kế, huấn luyện, kiểm thử. Phân định rõ chủ thể chịu trách nhiệm giải trình đối với các quyết định do hệ thống tạo ra.

3. Thúc đẩy hạnh phúc, thịnh vượng và sự phát triển bền vững của con người, cộng đồng và xã hội.

a) Lợi ích xã hội: Tổ chức, cá nhân xác định rõ lợi ích công cộng, giá trị gia tăng và tác động tích cực của hệ thống đối với con người và cộng đồng; có phương án xử lý, khắc phục tác động tiêu cực trước khi triển khai.

b) Phát triển bao trùm: Tổ chức, cá nhân đảm bảo ưu tiên thiết kế giao diện dễ tiếp cận, dễ sử dụng; thu hẹp khoảng cách số giữa các vùng miền, nhóm dân cư.

c) Phát triển bền vững: Tổ chức, cá nhân phát triển hoặc triển khai hệ thống trí tuệ nhân tạo có trách nhiệm xem xét mức tiêu thụ năng lượng, tài nguyên tính toán và tác động môi trường trong suốt vòng đời hệ thống; ưu tiên lựa chọn giải pháp kỹ thuật, hạ tầng và quy trình vận hành tiết kiệm năng lượng, hạn chế phát thải.

d) Tôn trọng văn hóa và giá trị xã hội: Tổ chức, cá nhân thiết kế hệ thống trí tuệ nhân tạo theo hướng phù hợp chuẩn mực đạo đức xã hội và bản sắc văn hóa Việt Nam; không được tạo ra các nội dung kỳ thị, phân biệt đối xử hoặc ảnh hưởng đến lợi ích cộng đồng.

4. Khuyến khích đổi mới sáng tạo và trách nhiệm xã hội trong nghiên cứu, phát triển và ứng dụng trí tuệ nhân tạo.

a) Khuyến khích đổi mới: Tổ chức, cá nhân triển khai thử nghiệm, thí điểm và mở rộng ứng dụng trí tuệ nhân tạo theo hướng có trách nhiệm; thúc đẩy nghiên cứu mở, chia sẻ tri thức phù hợp quy định pháp luật và bảo vệ quyền sở hữu trí tuệ.

b) Trách nhiệm xã hội: Tổ chức, cá nhân phát triển, triển khai và sử dụng hệ thống trí tuệ nhân tạo phân định rõ trách nhiệm của các chủ thể trong vòng đời hệ thống; bảo đảm có đầu mối tiếp nhận, xử lý khiếu nại và khắc phục hậu quả.

c) Nâng cao năng lực và hợp tác: Tổ chức, cá nhân chú trọng đào tạo về nhận thức, rủi ro đạo đức, kỹ năng sử dụng trí tuệ nhân tạo an toàn cho cán bộ, người lao động; tăng cường hợp tác và học hỏi các tiêu chuẩn quốc tế về đạo đức trí tuệ nhân tạo.

d) Hợp tác và học hỏi: Tổ chức, cá nhân tăng cường hợp tác quốc tế, tham gia sáng kiến, tiêu chuẩn, bộ quy tắc ứng xử về đạo đức trí tuệ nhân tạo; tận dụng sáng kiến khu vực tư nhân phục vụ lợi ích công theo điều kiện Việt Nam.

Điều 4. Hiệu lực thi hành

Thông tư này có hiệu lực thi hành kể từ ngày 10 tháng 3 năm 2026.

Điều 5. Trách nhiệm thi hành

1. Khung đạo đức trí tuệ nhân tạo quốc gia được rà soát, cập nhật định kỳ 03 năm một lần hoặc khi có thay đổi lớn về công nghệ, pháp luật và thực tiễn quản lý.

2. Bộ Khoa học và Công nghệ chủ trì tổ chức triển khai, thực hiện Thông tư này; ban hành hướng dẫn quản trị và đánh giá tuân thủ; cập nhật Khung đạo đức trí tuệ nhân tạo quốc gia phù hợp với yêu cầu phát triển và chuẩn mực quốc tế; xây dựng, vận hành nền tảng số và công cụ trắc nghiệm tự động để hỗ trợ các bộ, ngành, địa phương và doanh nghiệp tự đánh giá mức độ tuân thủ đạo đức trí tuệ nhân tạo.

3. Trong quá trình thực hiện, nếu phát sinh khó khăn, vướng mắc, đề nghị các cơ quan, tổ chức, cá nhân phản ánh về Bộ Khoa học và Công nghệ để nghiên cứu, sửa đổi, bổ sung cho phù hợp./.

Nơi nhận:
- Thủ tướng, các Phó Thủ tướng Chính phủ;
- Các Bộ, cơ quan ngang Bộ, cơ quan thuộc Chính phủ;
- Văn phòng TƯ Đảng và các Ban của Đảng;
- Văn phòng Tổng Bí thư;
- Văn phòng Quốc hội;
- Văn phòng Chủ tịch nước;
- Viện Kiểm sát nhân dân tối cao;
- Tòa án nhân dân tối cao;
- Kiểm toán nhà nước;
- Ủy ban Trung ương Mặt trận Tổ quốc Việt Nam;
- Cơ quan Trung ương của các tổ chức chính trị-xã hội;
- HĐND, UBND các tỉnh, thành phố;
- Sở Khoa học và Công nghệ các tỉnh, thành phố;
- Cục Kiểm tra văn bản QPPL, Bộ Tư pháp;
- Cổng thông tin điện tử Chính phủ; Công báo;
- Bộ KH&CN: Bộ trưởng, các Thứ trưởng; các cơ quan, đơn vị thuộc Bộ KH&CN, Cổng TTĐT Bộ KH&CN;
- Lưu: VT, CNS&CĐS.

BỘ TRƯỞNG

BỘ KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Mạnh Hùng

PHỤ LỤC I
HƯỚNG DẪN SỬ DỤNG KHUNG ĐẠO ĐỨC TRÍ TUỆ NHÂN TẠO QUỐC GIA THEO NHÓM CHỦ THỂ

(Ban hành kèm theo Thông tư số /2026/TT-BKHCN ngày tháng 03 năm 2026 của Bộ trưởng Bộ Khoa học và Công nghệ)

1. Đối với người sử dụng hệ thống trí tuệ nhân tạo

a) Xác định rõ nhu cầu, mục tiêu sử dụng và bối cảnh; sử dụng trí tuệ nhân tạo như công cụ hỗ trợ, không thay thế hoàn toàn phán đoán của con người đối với quyết định quan trọng.

b) Bảo vệ dữ liệu cá nhân, bí mật đời tư, thông tin mật và bí mật kinh doanh; không nhập hoặc chia sẻ thông tin nhạy cảm vào hệ thống trí tuệ nhân tạo khi chưa được phép hoặc chưa có biện pháp kiểm soát phù hợp.

c) Kiểm chứng thông tin, khuyến nghị hoặc kết quả đầu ra; đặc biệt đối với nội dung liên quan đến sức khỏe, tài chính, pháp lý, danh dự, nhân phẩm hoặc đời sống tinh thần của cá nhân.

d) Tôn trọng quyền và lợi ích hợp pháp của người khác; không sử dụng trí tuệ nhân tạo để tạo, phát tán nội dung sai sự thật, thao túng, phân biệt đối xử, xâm phạm quyền riêng tư, quyền sở hữu trí tuệ hoặc vi phạm pháp luật.

đ) Sử dụng các kênh phản ánh/khiếu nại khi phát hiện dấu hiệu thiên lệch, phân biệt đối xử, nội dung gây hại hoặc hành vi lạm dụng trí tuệ nhân tạo.

e) Tự nâng cao năng lực sử dụng trí tuệ nhân tạo an toàn và có trách nhiệm; cập nhật kiến thức về giới hạn của hệ thống, rủi ro đạo đức và quy định pháp luật có liên quan.

2. Đối với nhà phát triển, nhà cung cấp, bên triển khai hệ thống trí tuệ nhân tạo

a) Lồng ghép cân nhắc đạo đức trong toàn bộ vòng đời, bao gồm xác định mục tiêu, thiết kế, dữ liệu, huấn luyện, kiểm thử, triển khai, vận hành và cải tiến; ưu tiên giải pháp phòng ngừa vấn đề đạo đức từ sớm.

b) Bảo đảm minh bạch phù hợp: cung cấp thông tin ở mức hợp lý về mục tiêu, phạm vi, giới hạn của hệ thống; không thổi phồng năng lực; hỗ trợ người sử dụng nhận biết khi nào có sử dụng trí tuệ nhân tạo theo quy định của pháp luật.

c) Chủ động nhận diện và giảm thiểu thiên lệch; bảo đảm xem xét tác động đến nhóm dễ bị tổn thương; bảo đảm đa dạng hóa dữ liệu và nhóm phát triển.

d) Bảo vệ quyền riêng tư, dữ liệu cá nhân và an ninh thông tin mạng trong quá trình phát triển, cung cấp và triển khai; áp dụng biện pháp kiểm soát truy cập, lưu vết và quản lý vòng đời dữ liệu phù hợp.

đ) Thiết lập cơ chế giải trình và tiếp nhận phản ánh; phân định rõ trách nhiệm của các bên trong chuỗi cung ứng, bao gồm bên thứ ba, nhà thầu, đối tác; có cơ chế phối hợp khắc phục khi phát sinh vấn đề đạo đức.

e) Thực hiện ban hành Bộ quy tắc đạo đức trí tuệ nhân tạo nội bộ/theo ngành, lĩnh vực.

3. Đối với cơ quan nhà nước, tổ chức cung cấp dịch vụ công

a) Lồng ghép Khung đạo đức trong quá trình xây dựng, mua sắm, thuê dịch vụ, thí điểm và triển khai hệ thống trí tuệ nhân tạo; bảo đảm hiệu quả, minh bạch và không phát sinh thủ tục không cần thiết.

b) Bảo đảm quyết định cuối cùng thuộc thẩm quyền của con người theo quy định của pháp luật; hệ thống trí tuệ nhân tạo không thay thế trách nhiệm của người ra quyết định.

c) Thực hiện công khai, minh bạch phù hợp; tạo điều kiện để người dân được biết, được giải thích ở mức hợp lý và có cơ chế phản ánh, khiếu nại.

d) Bảo đảm tham vấn các bên liên quan, nhất là nhóm dễ bị tổn thương; chú trọng tác động xã hội và bảo đảm tiếp cận bao trùm.

đ) Thực hiện đánh giá và theo dõi tác động đạo đức trong quá trình vận hành, cải tiến; tích hợp với các báo cáo, hồ sơ hiện có để tránh trùng lặp.

PHỤ LỤC II
HƯỚNG DẪN MẪU PHIẾU TỰ ĐÁNH GIÁ TUÂN THỦ KHUNG ĐẠO ĐỨC TRÍ TUỆ NHÂN TẠO QUỐC GIA

(Kèm theo Thông tư số /2026/TT-BKHCN ngày tháng năm 2026 của Bộ trưởng Bộ Khoa học và Công nghệ)

Nội dung đánh giá	**Mức độ đáp ứng** (Đạt/Chưa đạt/Chưa áp dụng)	**Minh chứng**	**Hành động khắc phục** (nếu có)	**Đơn vị/cá nhân phụ trách & thời hạn**
A. Thông tin chung về hệ thống trí tuệ nhân tạo (tên, mục tiêu, phạm vi, người bị ảnh hưởng, giai đoạn vòng đời)				
B. Nguyên tắc 1: An toàn, độ tin cậy, không gây hại				
1. Có cơ chế kiểm thử, xác nhận, giám sát chất lượng đầu ra phù hợp				
2. Có cơ chế kiểm soát/can thiệp của con người đối với trường hợp quan trọng				
3. Có quy trình tiếp nhận phản ánh và xử lý sự cố liên quan tác động tiêu cực				
C. Nguyên tắc 2: Quyền con người, công bằng, minh bạch, không phân biệt đối xử				
1. Có biện pháp nhận diện và giảm thiểu thiên lệch dữ liệu/mô hình				

Nội dung đánh giá	**Mức độ đáp ứng** (Đạt/Chưa đạt/Chưa áp dụng)	**Minh chứng**	**Hành động khắc phục** (nếu có)	**Đơn vị/cá nhân phụ trách & thời hạn**
2. Có thông báo minh bạch phù hợp tới người sử dụng/người bị ảnh hưởng				
3. Có cơ chế giải thích ở mức hợp lý và lưu vết quyết định				
D. Nguyên tắc 3: Hạnh phúc, thịnh vượng, phát triển bền vững				
1. Mục tiêu/giá trị gia tăng xã hội của hệ thống được xác định và theo dõi				
2. Có cân nhắc tác động bao trùm và giảm khoảng cách số				
3. Có cân nhắc tác động môi trường và sử dụng tài nguyên tính toán hợp lý				
E. Nguyên tắc 4: Đổi mới sáng tạo và trách nhiệm xã hội				
1. Trách nhiệm và đầu mối giải trình được phân công rõ ràng				
2. Có cơ chế đào tạo/nâng cao nhận thức cho nhân sự và người sử dụng				
3. Có cơ chế tham vấn các bên liên quan khi cần thiết				

Nội dung đánh giá	**Mức độ đáp ứng** (Đạt/Chưa đạt/Chưa áp dụng)	**Minh chứng**	**Hành động khắc phục** (nếu có)	**Đơn vị/cá nhân phụ trách & thời hạn**
G. Kết luận và kiến nghị (duy trì/điều chỉnh/tạm dừng/mở rộng triển khai)				