BẢN TIN
CHIẾN LƯỢC PHÁT TRIỂN

KHOA HỌC CÔNG NGHỆ KINH TẾ

Số 5
2024

(BẢN TIN CHỌN LỌC PHỤC VỤ LÃNH ĐẠO)

KHUYẾN NGHỊ CỦA LIÊN HỢP QUỐC VỀ ĐẠO ĐỨC, AN TOÀN VÀ TRÁCH NHIỆM TRONG PHÁT TRIỂN VÀ ỨNG DỤNG TRÍ TUỆ NHÂN TẠO

NASATI

BỘ KHOA HỌC VÀ CÔNG NGHỆ
CỤC THÔNG TIN KHOA HỌC VÀ CÔNG NGHỆ QUỐC GIA

CỤC THÔNG TIN KHOA HỌC VÀ CÔNG NGHỆ QUỐC GIA

Địa chỉ: 24, Lý Thường Kiệt, Hoàn Kiếm, Hà Nội.

Tel: (024)38262718, Fax: (024)39349127

BAN BIÊN TẬP

TS. Trần Đắc Hiến (*Trưởng ban*);

ThS. Nguyễn Lê Hằng; ThS. Phùng Anh Tiến.

MỤC LỤC

KHUYẾN NGHỊ CỦA LIÊN HỢP QUỐC VỀ ĐẠO ĐỨC, AN TOÀN
VÀ TRÁCH NHIỆM TRONG PHÁT TRIỂN VÀ ỨNG DỤNG TRÍ TUỆ NHÂN TẠO

KHUYẾN NGHỊ CỦA LIÊN HỢP QUỐC VỀ ĐẠO ĐỨC, AN TOÀN VÀ TRÁCH NHIỆM TRONG PHÁT TRIỂN VÀ ỨNG DỤNG TRÍ TUỆ NHÂN TẠO

Giới thiệu

Thời gian qua, trí tuệ nhân tạo (AI) đã có bước phát triển nhanh và đột phá ở trên thế giới và Việt Nam, tuy nhiên, việc quản lý AI, phát triển các sản phẩm AI có đạo đức, có trách nhiệm là vấn đề đang được các quốc gia, các tổ chức quốc tế quan tâm, tham gia tìm phương án giải quyết, gồm cả Liên hợp quốc, trong đó có vai trò đặc biệt của Hoa Kỳ. Tại Hoa Kỳ, chính quyền liên bang đang thúc đẩy các nhà lập pháp hoàn thiện hành lang pháp lý về AI trong bối cảnh ngày càng có nhiều tiếng nói cảnh báo về những rủi ro trong việc sử dụng công nghệ này. Tháng 10/2023, Tổng thống Hoa Kỳ Joe Biden đã ban hành Sắc lệnh về phát triển và sử dụng AI an toàn và đáng tin cậy của Hoa Kỳ, nhằm giảm thiểu những rủi ro mà công nghệ AI có thể gây ra, thiết lập các tiêu chuẩn mới về an toàn và bảo mật, bảo vệ quyền riêng tư của người dùng, thúc đẩy đổi mới và cạnh tranh trong lĩnh vực AI. Tháng 11/2023, Hoa Kỳ, Anh và hơn 10 quốc gia khác đã công bố một thỏa thuận quốc tế chi tiết đầu tiên về cách thức bảo vệ an toàn cho AI, đồng thời thúc đẩy các công ty tạo ra các hệ thống AI "an toàn ngay từ trong thiết kế". Ngày 13/3/2024 vừa qua, Nghị viện châu Âu đã thông qua Luật Trí tuệ nhân tạo đầu tiên trên thế giới.

Gần đây nhất, lần đầu tiên, Liên hợp quốc đưa AI vào thảo luận để đưa ra nghị quyết với các khuyến nghị sử dụng AI an toàn và có trách nhiệm. Ngày 21/3/2024, Đại hội đồng Liên hợp quốc đã thông qua nghị quyết toàn cầu đầu tiên về AI nhằm thúc đẩy các hệ thống AI "an toàn, bảo mật và đáng tin cậy", mang lại lợi ích phát triển bền vững cho tất cả mọi người, kêu gọi các nước chung tay bảo vệ quyền con người và kiểm soát những rủi ro tiềm ẩn từ công nghệ này. Đây là lần đầu tiên Hội đồng thông qua một nghị quyết về quản lý lĩnh vực mới nổi. Nghị quyết không mang tính ràng buộc pháp lý do Hoa Kỳ đề xuất và được hơn 120 quốc gia khác đồng bảo trợ. Nghị quyết được thông qua với sự đồng thuận của toàn bộ 193 nước thành viên Liên hợp quốc. Trước đó, ngày 23/11/2021, các quốc gia thành viên của Tổ chức Giáo dục, Khoa học và văn hóa của Liên hợp quốc (UNESCO) đã thông qua Khuyến nghị về Đạo đức AI.

Bản tin này sẽ giới thiệu những nội dung chính của Nghị quyết toàn cầu đầu tiên về AI và Khuyến nghị về Đạo đức AI của Liên hợp quốc.

1. NGHỊ QUYẾT CỦA LIÊN HỢP QUỐC VỀ TRÍ TUỆ NHÂN TẠO

Ngày 21/3/2024, Đại hội đồng Liên hợp quốc đã thông qua Nghị quyết toàn cầu đầu tiên về AI - "Nắm bắt cơ hội của các hệ thống AI an toàn, bảo mật và đáng tin cậy để phát triển bền vững". Đây là một nghị quyết mang tính bước ngoặt về việc thúc đẩy các hệ thống AI "an toàn, bảo mật và đáng tin cậy". Dự thảo Nghị quyết do Hoa Kỳ dẫn đầu được thông qua nhấn mạnh sự tôn trọng, bảo vệ và thúc đẩy nhân quyền trong thiết kế, phát triển, triển khai và sử dụng AI; công nhận tiềm năng của các hệ thống AI trong việc tăng tốc và tạo điều kiện thuận lợi cho tiến trình đạt được 17 Mục tiêu Phát triển Bền vững.

Đại sứ Hoa Kỳ tại Liên hợp quốc Linda Thomas - Greenfield nhấn mạnh: "Tất cả các quốc gia đã có cùng tiếng nói trong vấn đề này và cùng nhau lựa chọn quản lý AI thay vì để nó chi phối chúng ta". Nghị quyết là sáng kiến mới nhất trong một loạt sáng kiến của các chính phủ trên thế giới nhằm định hình sự phát triển của AI, trong bối cảnh lo ngại việc nó có thể được sử dụng để phá vỡ các quy trình dân chủ, làm gia tăng các hành vị gian lận hoặc gây ra tình trạng mất việc làm nghiêm trọng, cùng nhiều tác hại khác. Theo Nghị quyết, việc thiết kế, phát triển, triển khai và sử dụng hệ thống AI sai mục đích hoặc có ác ý ẩn chứa nhiều rủi ro làm suy yếu việc bảo vệ, thúc đẩy và hưởng thụ các quyền con người và các quyền tự do cơ bản.

1.1. Những nội dung được thông qua trong Nghị quyết

- Các hệ thống AI an toàn, bảo mật và đáng tin cậy đề cập đến các hệ thống AI trong lĩnh vực phi quân sự, có vòng đời bao gồm các giai đoạn: tiền thiết kế, thiết kế, phát triển, đánh giá, thử nghiệm, triển khai, sử dụng, bán, mua sắm, vận hành và ngừng hoạt động phải lấy con người làm trung tâm, đáng tin cậy, có thể giải thích được, có đạo đức, toàn diện, tôn trọng đầy đủ, thúc đẩy và bảo vệ nhân quyền và luật pháp quốc tế, bảo vệ quyền riêng tư, định hướng phát triển bền vững và có trách nhiệm – có khả năng đẩy nhanh và tạo điều kiện cho tiến trình đạt được tất cả 17 Mục tiêu Phát triển Bền vững và phát triển bền vững ở ba khía cạnh – kinh tế, xã hội và môi trường – một cách cân bằng và tổng hợp; thúc đẩy chuyển đổi số; thúc đẩy hòa bình; vượt qua khoảng cách kỹ thuật số giữa các quốc gia; và thúc đẩy và bảo vệ việc thụ hưởng các quyền con người và các quyền tự do cơ bản cho tất cả mọi người, đồng thời lấy con người làm trung tâm.

- Việc thiết kế, phát triển, triển khai và sử dụng hệ thống AI không đúng hoặc có ác ý, chẳng hạn như không có biện pháp bảo vệ đầy đủ hoặc theo cách không phù hợp với luật pháp quốc tế, sẽ gây ra những rủi ro có thể cản trở tiến trình hướng tới đạt được Chương trình nghị sự 2030 về Phát triển bền vững và các Mục tiêu Phát triển Bền vững và làm suy yếu sự phát triển bền vững ở ba khía cạnh – kinh tế, xã hội và môi trường; mở rộng khoảng cách số giữa các quốc gia; củng cố sự bất bình đẳng và thành kiến về mặt cấu trúc; dẫn đến sự phân biệt đối xử; làm suy yếu tính toàn vẹn thông tin và quyền truy cập thông tin; cắt giảm việc bảo vệ, thúc đẩy và thụ hưởng các quyền con người và các quyền tự do cơ bản, bao gồm cả quyền không bị can thiệp bất hợp pháp hoặc tùy

tiện vào quyền riêng tư của một người; và tăng nguy cơ xảy ra tai nạn cũng như các mối đe dọa phức tạp từ các tác nhân độc hại.

- Sự tăng tốc nhanh chóng của việc thiết kế, phát triển, triển khai và sử dụng các hệ thống AI cũng như sự thay đổi công nghệ nhanh chóng và tác động tiềm tàng của chúng trong việc đẩy nhanh việc đạt được các Mục tiêu Phát triển Bền vững, do đó nhấn mạnh tính cấp bách của việc đạt được sự đồng thuận toàn cầu về an toàn, bảo mật và hệ thống AI đáng tin cậy; tạo điều kiện hợp tác quốc tế toàn diện nhằm xây dựng và sử dụng các biện pháp bảo vệ, thông lệ và tiêu chuẩn hiệu quả, có khả năng tương tác quốc tế nhằm thúc đẩy đổi mới và ngăn chặn sự phân tán trong quản lý các hệ thống AI an toàn, bảo mật và đáng tin cậy; và cũng thừa nhận AI hiện có và các khoảng cách số khác cũng như mức độ phát triển công nghệ khác nhau giữa và trong các quốc gia, rằng các nước đang phát triển phải đối mặt với những thách thức đặc biệt trong việc theo kịp tốc độ tăng tốc nhanh chóng này, vốn gây trở ngại cho sự phát triển bền vững, nhu cầu thu hẹp phạm vi hiện có sự chênh lệch giữa các nước phát triển và đang phát triển về điều kiện, khả năng và năng lực, do đó cũng nhấn mạnh tính cấp thiết của việc tăng cường xây dựng năng lực và hỗ trợ kỹ thuật và tài chính cho các nước đang phát triển để thu hẹp khoảng cách kỹ thuật số giữa và trong các nước và hỗ trợ các nước đang phát triển một cách hiệu quả, công bằng và có ý nghĩa, giúp họ có cơ hội tham gia và đại diện trong các tiến trình và diễn đàn quốc tế về quản trị hệ thống AI,

- Quản trị hệ thống AI là một lĩnh vực đang phát triển và cần tiếp tục thảo luận về các phương pháp quản trị khả thi phù hợp, dựa trên luật pháp quốc tế, có thể tương tác, linh hoạt, thích ứng, toàn diện, đáp ứng các nhu cầu và năng lực khác nhau của các quốc gia phát triển và các nước đang phát triển vì lợi ích của tất cả mọi người.

1.2. Những nội dung quan trọng được nhấn mạnh trong Nghị quyết

1. Quyết tâm thu hẹp khoảng cách AI và các khoảng cách kỹ thuật số giữa và trong các quốc gia.

2. Quyết tâm thúc đẩy các hệ thống AI an toàn, bảo mật và đáng tin cậy nhằm đẩy nhanh tiến độ hướng tới thực hiện đầy đủ Chương trình nghị sự 2030 về Phát triển bền vững, tiếp tục thu hẹp khoảng cách về AI và các khoảng cách kỹ thuật số khác giữa và trong các quốc gia; và nhấn mạnh sự cần thiết của tiêu chuẩn về hệ thống AI an toàn, bảo mật và đáng tin cậy để thúc đẩy, không cản trở, chuyển đổi kỹ thuật số và tiếp cận công bằng các lợi ích của chúng nhằm đạt được tất cả 17 Mục tiêu Phát triển Bền vững và phát triển bền vững ở ba khía cạnh – kinh tế, xã hội và môi trường – và giải quyết các thách thức toàn cầu chung khác, đặc biệt đối với các nước đang phát triển.

3. Khuyến khích Quốc gia Thành viên và mời các bên liên quan từ mọi khu vực và quốc gia, bao gồm cả khu vực tư nhân, tổ chức quốc tế và khu vực, xã hội dân sự, giới truyền thông và tổ chức nghiên cứu, để phát triển và hỗ trợ các phương pháp tiếp cận và khuôn khổ pháp lý và quản trị liên quan đến các hệ thống AI an toàn, bảo mật và đáng tin cậy. Điều này nhằm tạo ra một hệ sinh thái thuận lợi ở mọi cấp độ, kích thích đổi mới, khởi nghiệp và chia sẻ kiến thức và công nghệ.

4. Kêu gọi các Quốc gia Thành viên và mời các bên liên quan khác hành động để hợp tác và hỗ trợ các nước đang phát triển hướng tới khả năng tiếp cận toàn diện và công bằng các lợi ích của chuyển đổi số cũng như các hệ thống AI an toàn, bảo mật và đáng tin cậy, bằng cách:

(a) Mở rộng sự tham gia của tất cả các quốc gia, đặc biệt là các nước đang phát triển, vào chuyển đổi số để khai thác lợi ích và tham gia hiệu quả vào việc phát triển, triển khai và sử dụng các hệ thống AI an toàn, bảo mật và đáng tin cậy, bao gồm cả việc xây dựng năng lực liên quan đến hệ thống AI; việc thúc đẩy các hoạt động chia sẻ kiến thức và chuyển giao công nghệ là một khía cạnh quan trọng của việc xây dựng năng lực, nhấn mạnh sự cần thiết phải thu hẹp khoảng cách AI và các khoảng cách số khác;

(b) Tăng cường kết nối cơ sở hạ tầng kỹ thuật số và khả năng tiếp cận đổi mới công nghệ thông qua quan hệ đối tác mạnh mẽ hơn để giúp các nước đang phát triển tham gia hiệu quả trong suốt vòng đời của hệ thống AI và đẩy nhanh sự đóng góp toàn diện và tích cực của hệ thống AI cho xã hội, bao gồm cả việc hướng tới hiện thực hóa đầy đủ các mục tiêu Chương trình nghị sự 2030 và các Mục tiêu Phát triển Bền vững, đồng thời bảo đảm rằng các hệ thống AI trên toàn thế giới được an toàn, bảo mật và đáng tin cậy trong suốt vòng đời của chúng;

(c) Nâng cao năng lực của các nước đang phát triển, đặc biệt là các nước kém phát triển nhất, trong việc giải quyết các trở ngại lớn trong việc tiếp cận lợi ích của các công nghệ mới và mới nổi cũng như đổi mới AI để đạt được tất cả 17 Mục tiêu Phát triển Bền vững, bao gồm cả việc mở rộng quy mô việc sử dụng các nguồn khoa học, công nghệ, nghiên cứu và phát triển có giá cả phải chăng, bao gồm cả thông qua tăng cường quan hệ đối tác;

(d) Tăng nguồn tài trợ cho nghiên cứu và đổi mới liên quan đến các Mục tiêu Phát triển Bền vững, bao gồm công nghệ số và hệ thống AI an toàn, bảo mật và đáng tin cậy, đồng thời xây dựng năng lực ở tất cả các khu vực và quốc gia để đóng góp và hưởng lợi từ các nghiên cứu và đổi mới này;

(e) Tạo điều kiện cho môi trường đổi mới quốc tế nhằm nâng cao khả năng của các nước đang phát triển trong việc phát triển chuyên môn và năng lực kỹ thuật, khai thác dữ liệu và tài nguyên máy tính cũng như các phương pháp tiếp cận, khuôn khổ và năng lực quản lý, đồng thời tạo ra một môi trường thuận lợi toàn diện ở tất cả các cấp cho các giải pháp dựa trên hệ thống AI an toàn, bảo mật và đáng tin cậy;

(f) Huy động khẩn cấp các phương tiện thực hiện như chuyển giao công nghệ, xây dựng năng lực để thu hẹp khoảng cách AI và các khoảng cách số khác, hỗ trợ kỹ thuật và tài chính cho các nước đang phát triển liên quan đến hệ thống AI phù hợp với nhu cầu quốc gia, chính sách và ưu tiên của các nước đang phát triển;

(g) Thúc đẩy khả năng tiếp cận và thiết kế, phát triển, triển khai và sử dụng các hệ thống AI an toàn, bảo mật và đáng tin cậy để đạt được sự phát triển bền vững ở ba khía cạnh – kinh tế, xã hội và môi trường.

5. Nhân quyền và các quyền tự do cơ bản phải được tôn trọng, bảo vệ và thúc đẩy trong suốt vòng đời của hệ thống AI, Liên hợp quốc kêu gọi tất cả các Quốc gia Thành viên và, nếu có, các bên liên quan khác kiềm chế hoặc ngừng sử dụng các hệ thống AI không tuân thủ luật nhân quyền quốc tế hoặc gây ra rủi ro không đáng có cho việc thụ hưởng nhân quyền, đặc biệt là đối với những người ở trong hoàn cảnh dễ bị tổn thương và tái khẳng định rằng các quyền tương tự mà mọi người có được ngoài đời thực cũng phải được bảo vệ trực tuyến, bao gồm cả trong suốt vòng đời của hệ thống AI.

6. Khuyến khích tất cả các Quốc gia Thành viên, khi thích hợp, phù hợp với các ưu tiên và hoàn cảnh quốc gia của họ, đồng thời thực hiện các khuôn khổ và cách tiếp cận quản lý quốc gia riêng biệt của họ, thúc đẩy các hệ thống AI an toàn, bảo mật và đáng tin cậy một cách toàn diện và công bằng, vì lợi ích của tất cả mọi người; đồng thời thúc đẩy môi trường thuận lợi cho các hệ thống như vậy giải quyết những thách thức lớn nhất của thế giới, bao gồm đạt được sự phát triển bền vững ở ba khía cạnh - kinh tế, xã hội và môi trường - với sự quan tâm cụ thể đến các nước đang phát triển:

(a) Thúc đẩy việc phát triển và thực hiện các khuôn khổ và cách tiếp cận quy định và quản trị trong nước để hỗ trợ đầu tư và đổi mới AI có trách nhiệm và toàn diện cho phát triển bền vững phát triển, đồng thời thúc đẩy các hệ thống AI an toàn, bảo mật và đáng tin cậy;

(b) Khuyến khích các biện pháp hiệu quả, thúc đẩy đổi mới để nhận dạng, phân loại, đánh giá, thử nghiệm, ngăn ngừa và giảm thiểu các lỗ hổng và rủi ro trong quá trình thiết kế và phát triển cũng như trước khi triển khai và sử dụng hệ thống AI;

(c) Khuyến khích việc kết hợp các cơ chế phản hồi để cho phép người dùng cuối và bên thứ ba phát hiện và báo cáo dựa trên bằng chứng về các lỗ hổng kỹ thuật và, nếu phù hợp, lạm dụng hệ thống AI và các sự cố AI sau quá trình phát triển, thử nghiệm và triển khai để giải quyết chúng;

(d) Nâng cao nhận thức và hiểu biết của công chúng về các chức năng, năng lực, hạn chế và lĩnh vực cốt lõi của việc sử dụng dân sự phù hợp các hệ thống AI;

(e) Thúc đẩy việc phát triển, triển khai và công bố các cơ chế giám sát và quản lý rủi ro, cơ chế bảo mật dữ liệu, bao gồm các chính sách về quyền riêng tư và bảo vệ dữ liệu cá nhân, cũng như đánh giá tác động nếu phù hợp, trong suốt vòng đời của hệ thống AI;

(f) Tăng cường đầu tư vào việc phát triển và thực hiện các biện pháp bảo vệ hiệu quả, bao gồm an ninh vật lý, an ninh hệ thống AI và quản lý rủi ro trong suốt vòng đời của hệ thống AI;

(g) Khuyến khích phát triển và triển khai các công cụ, tiêu chuẩn hoặc thông lệ kỹ thuật hiệu quả, có thể truy cập, thích ứng và có khả năng tương tác quốc tế, bao gồm các cơ chế xác thực nội dung và xuất xứ đáng tin cậy - chẳng hạn như hình mờ hoặc gắn nhãn, khi khả thi và phù hợp về mặt kỹ thuật, cho phép người dùng nhận dạng thông tin, phân biệt hoặc xác định nguồn gốc của nội dung kỹ thuật số;

(h) Tạo điều kiện cho việc phát triển và thực hiện các khuôn khổ, thông lệ và tiêu chuẩn hiệu quả, có khả năng tương tác quốc tế để đào tạo và thử nghiệm các hệ thống AI nhằm tăng cường hoạch định chính sách và giúp bảo vệ các cá nhân khỏi mọi hình thức phân biệt đối xử, thiên vị, lạm dụng hoặc các tác hại khác;

(i) Khuyến khích việc thực hiện các biện pháp bảo vệ thích hợp để tôn trọng quyền sở hữu trí tuệ, bao gồm cả nội dung được bảo vệ bản quyền, đồng thời thúc đẩy đổi mới;

(j) Bảo vệ quyền riêng tư và bảo vệ dữ liệu cá nhân khi kiểm tra và đánh giá hệ thống cũng như các yêu cầu về tính minh bạch và báo cáo tuân thủ các khung pháp lý quốc tế, quốc gia và địa phương hiện hành, bao gồm cả việc sử dụng dữ liệu cá nhân trong suốt vòng đời của hệ thống AI;

(k) Thúc đẩy tính minh bạch, khả năng dự đoán, độ tin cậy và tính dễ hiểu trong suốt vòng đời của các hệ thống AI, bao gồm việc đưa ra thông báo và giải thích cũng như thúc đẩy sự giám sát của con người, chẳng hạn như thông qua việc xem xét các hệ thống tự động hóa các quyết định và quy trình liên quan hoặc các giải pháp thay thế đưa ra quyết định của con người hoặc biện pháp khắc phục hiệu quả và trách nhiệm giải trình đối với những người bị ảnh hưởng bất lợi bởi các quyết định tự động của hệ thống AI;

(l) Tăng cường đầu tư vào việc phát triển và thực hiện các biện pháp bảo vệ hiệu quả, bao gồm đánh giá rủi ro và tác động, trong suốt vòng đời của hệ thống AI để bảo vệ việc thực hiện và giảm thiểu tác động tiềm ẩn đối với việc thụ hưởng đầy đủ và hiệu quả các quyền con người và các quyền tự do cơ bản;

(m) Thúc đẩy các hệ thống AI nhằm nâng cao, bảo vệ và bảo tồn sự đa dạng về ngôn ngữ và văn hóa, có tính đến tính đa ngôn ngữ trong dữ liệu đào tạo và trong suốt vòng đời của hệ thống AI, đặc biệt đối với các mô hình ngôn ngữ lớn;

(n) Tăng cường chia sẻ thông tin theo các điều khoản được thống nhất giữa các thực thể trong suốt vòng đời của hệ thống AI để xác định, hiểu và hành động bằng cách sử dụng các phương pháp, chính sách và phương pháp tiếp cận tốt nhất dựa trên khoa học và dựa trên bằng chứng đối với hệ thống AI nhằm tối đa hóa lợi ích và giảm thiểu rủi ro tiềm ẩn trong vòng đời của hệ thống AI, bao gồm cả hệ thống AI tiên tiến;

(o) Khuyến khích nghiên cứu và hợp tác quốc tế để hiểu và cân bằng các lợi ích và giải quyết rủi ro tiềm ẩn liên quan đến hệ thống AI trong việc thu hẹp khoảng cách số và đạt được tất cả 17 Mục tiêu Phát triển Bền vững, bao gồm cả vai trò nhân rộng các giải pháp số như hệ thống AI nguồn mở;

(p) Kêu gọi các Quốc gia Thành viên áp dụng các biện pháp cụ thể để thu hẹp khoảng cách kỹ thuật số về giới và bảo đảm đặc biệt chú ý đến khả năng tiếp cận, khả năng chi trả, kiến thức kỹ thuật số, quyền riêng tư và an toàn trực tuyến, nhằm tăng cường sử dụng công nghệ kỹ thuật số, bao gồm cả hệ thống AI và lồng ghép các vấn đề về giới tính và bình đẳng chủng tộc vào các quyết định chính sách và khuôn khổ hướng dẫn;

(q) Khuyến khích nghiên cứu và hợp tác quốc tế để phát triển các biện pháp xác định và đánh giá tác động của việc triển khai hệ thống AI trên thị trường lao động và hỗ trợ giảm thiểu những hậu quả tiêu cực tiềm ẩn đối với lực lượng lao động, đặc biệt là ở các nước đang phát triển, các nước kém phát triển nhất và thúc đẩy các chương trình nhằm đào tạo kỹ thuật số, xây dựng năng lực, hỗ trợ đổi mới và tăng cường khả năng tiếp cận lợi ích của hệ thống AI.

7. Cũng thừa nhận rằng dữ liệu là nền tảng cho sự phát triển và vận hành các hệ thống AI; nhấn mạnh rằng quản trị dữ liệu công bằng, toàn diện, có trách nhiệm và hiệu quả, cải thiện việc tạo dữ liệu, khả năng tiếp cận và cơ sở hạ tầng cũng như việc sử dụng hàng hóa công cộng kỹ thuật số là điều cần thiết để khai thác tiềm năng của hệ thống AI an toàn, bảo mật và đáng tin cậy cho sự phát triển bền vững; kêu gọi các quốc gia thành viên chia sẻ các thực tiễn tốt nhất về quản trị dữ liệu và thúc đẩy hợp tác quốc tế, hợp tác và hỗ trợ về quản trị dữ liệu để có tính nhất quán và khả năng tương tác cao hơn, chia sẻ các phương pháp tiếp cận nhằm thúc đẩy các luồng dữ liệu xuyên biên giới đáng tin cậy cho các hệ thống AI an toàn, bảo mật và đáng tin cậy, và làm cho sự phát triển của nó trở nên toàn diện, công bằng, hiệu quả và có lợi hơn cho tất cả mọi người.

8. Thừa nhận tầm quan trọng của việc tiếp tục thảo luận về sự phát triển trong lĩnh vực quản trị AI để các phương pháp tiếp cận quốc tế theo kịp sự phát triển của hệ thống AI và việc sử dụng chúng; và khuyến khích cộng đồng quốc tế tiếp tục nỗ lực thúc đẩy nghiên cứu, lập bản đồ và phân tích toàn diện nhằm mang lại lợi ích cho tất cả các bên về các tác động và ứng dụng tiềm năng mà hệ thống AI và sự thay đổi công nghệ nhanh chóng có thể có trong việc phát triển các công nghệ hiện có cũng như thúc đẩy quá trình phát triển đạt được tất cả 17 Mục tiêu Phát triển Bền vững; cung cấp thông tin về cách phát triển, thúc đẩy và triển khai các biện pháp bảo vệ, thông lệ, tiêu chuẩn và công cụ hiệu quả, có khả năng tương tác quốc tế dành cho các nhà thiết kế, nhà phát triển, người đánh giá, người triển khai, người dùng và các bên liên quan khác về AI để bảo đảm an toàn, bảo mật và đáng tin cậy hệ thống AI; cũng như nhấn mạnh sự cần thiết của các Chính phủ, khu vực tư nhân, xã hội dân sự, các tổ chức quốc tế và khu vực, các viện nghiên cứu cũng như cộng đồng kỹ thuật và tất cả các bên liên quan khác tiếp tục hợp tác cùng nhau; cũng như thừa nhận sự cần thiết phải có sự tham gia gắn kết, hiệu quả, phối hợp và toàn diện hơn của tất cả các cộng đồng, đặc biệt là từ các nước đang phát triển, trong việc quản trị toàn diện các hệ thống AI an toàn, bảo mật và đáng tin cậy.

9. Khuyến khích khu vực tư nhân tuân thủ luật pháp quốc tế và trong nước hiện hành cũng như hành động phù hợp với Nguyên tắc Hướng dẫn của Liên Hợp Quốc về Kinh doanh và Nhân quyền: Thực hiện Khung “Bảo vệ, Tôn trọng và Khắc phục” của Liên Hợp Quốc; thừa nhận tầm quan trọng của việc tiếp cận toàn diện và công bằng hơn với các lợi ích của hệ thống AI an toàn, bảo mật và đáng tin cậy; và thừa nhận nhu cầu tăng cường hợp tác, bao gồm giữa và trong khu vực công và tư nhân cũng như xã hội dân sự, các viện nghiên cứu và cộng đồng kỹ thuật, để cung cấp và thúc đẩy môi trường kinh doanh, các hoạt động kinh tế và thương mại công bằng, cởi mở, toàn diện và không phân biệt đối xử, hệ sinh thái và thị trường cạnh tranh trong suốt vòng đời của AI an

toàn, bảo mật và đáng tin cậy; cũng như khuyến khích các quốc gia thành viên xây dựng các chính sách và quy định nhằm thúc đẩy cạnh tranh trong các hệ thống AI an toàn, bảo mật và đáng tin cậy cũng như các công nghệ liên quan, bao gồm việc hỗ trợ và tạo cơ hội mới cho các doanh nghiệp nhỏ và doanh nhân cũng như nhân tài kỹ thuật, đồng thời tạo điều kiện cạnh tranh công bằng trong thị trường AI, thông qua đầu tư, đặc biệt đối với các nước đang phát triển.

10. Kêu gọi các cơ quan, quỹ, chương trình chuyên môn, các đơn vị, cơ quan và văn phòng khác cũng như các tổ chức liên quan của hệ thống Liên Hợp Quốc, trong phạm vi nhiệm vụ và nguồn lực tương ứng của mình, tiếp tục đánh giá và tăng cường phản ứng của mình để tận dụng các cơ hội và giải quyết các thách thức được đặt ra bởi các hệ thống AI theo cách hợp tác, phối hợp và toàn diện, thông qua các cơ chế liên cơ quan thích hợp, bao gồm bằng cách tiến hành nghiên cứu, lập bản đồ và phân tích mang lại lợi ích cho tất cả các bên về các tác động và ứng dụng tiềm năng; báo cáo về tiến độ và thách thức trong việc giải quyết các vấn đề; và hợp tác và hỗ trợ các nước đang phát triển xây dựng năng lực, tiếp cận và chia sẻ lợi ích của các hệ thống AI an toàn, bảo mật và đáng tin cậy nhằm đạt được tất cả 17 Mục tiêu Phát triển Bền vững và phát triển bền vững ở ba khía cạnh – kinh tế, xã hội và môi trường; nhấn mạnh sự cần thiết phải đóng cửa AI và các khoảng cách kỹ thuật số khác giữa và trong các quốc gia.

11. Thừa nhận rằng hệ thống của Liên hợp quốc, nhất quán với nhiệm vụ của mình, góp phần đặc biệt vào việc đạt được sự đồng thuận toàn cầu về các hệ thống AI an toàn, bảo mật và đáng tin cậy, phù hợp với luật pháp quốc tế, đặc biệt là Hiến chương Liên Hợp Quốc; Tuyên ngôn Quốc tế Nhân quyền; và Chương trình nghị sự 2030 về Phát triển bền vững, bao gồm việc thúc đẩy hợp tác quốc tế toàn diện và tạo điều kiện cho sự tham gia, đại diện của các nước đang phát triển trong các cuộc thảo luận.

2. KHUYẾN NGHỊ VỀ ĐẠO ĐỨC TRÍ TUỆ NHÂN TẠO

Vào tháng 11 năm 2021, các quốc gia thành viên UNESCO đã thông qua Khuyến nghị về đạo đức AI, một quy chuẩn đầu tiên về đạo đức AI toàn cầu. Văn bản lịch sử này, trong khía cạnh phổ quát của nó, xác định các giá trị và nguyên tắc chung sẽ hướng dẫn việc hình thành cơ sở hạ tầng pháp lý và chính sách nhằm bảo đảm sự phát triển và sử dụng AI có đạo đức và có trách nhiệm, phản đối việc sử dụng công nghệ này cho các ứng dụng "xâm hại", "vi phạm quyền con người và các quyền tự do cơ bản".

Theo UNESCO, hướng dẫn này đóng vai trò như bộ "khuyến nghị" toàn cầu thay vì một thỏa thuận ràng buộc. Hướng dẫn kêu gọi "minh bạch hơn trong việc kiểm soát dữ liệu cá nhân" và "nhận thức rõ hơn về khả năng AI bắt chước đặc điểm, hành vi của con người". UNESCO muốn bảo đảm rằng "sự thống trị của ngôn ngữ tiếng Anh trong AI không gây bất lợi cho các ngôn ngữ thiểu số, sự đa dạng và quan điểm văn hóa rộng lớn hơn". UNESCO cũng cảnh báo "việc tương tác liên tục với công nghệ AI, bao gồm thông qua các thuật toán mạng xã hội, có thể tác động tiêu cực đến sức khỏe tinh thần của cả trẻ em và người lớn".

2.1. Mục tiêu của Khuyến nghị

Mục tiêu tổng quát

Khuyến nghị này cung cấp cơ sở để giúp các hệ thống AI hoạt động vì lợi ích của nhân loại, cá nhân, xã hội, môi trường và hệ sinh thái, đồng thời ngăn ngừa tác hại; khuyến khích việc sử dụng các hệ thống AI một cách hòa bình.

Ngoài các khuôn khổ đạo đức hiện có liên quan đến AI trên toàn thế giới, Khuyến nghị này cũng nhằm mang đến một công cụ quy chuẩn được chấp nhận trên toàn cầu, không chỉ tập trung vào việc trình bày rõ ràng các giá trị và nguyên tắc mà còn vào việc hiện thực hóa chúng trên thực tế, thông qua các khuyến nghị chính sách cụ thể, nhấn mạnh vào các vấn đề hội nhập, bình đẳng giới, bảo vệ môi trường và hệ sinh thái.

Do sự phức tạp của các vấn đề đạo đức xung quanh AI đòi hỏi sự hợp tác của nhiều bên liên quan ở các cấp độ và lĩnh vực khác nhau của cộng đồng quốc tế, khu vực và quốc gia, Khuyến nghị này cho phép các bên liên quan chia sẻ trách nhiệm dựa trên cuộc đối thoại toàn cầu và liên văn hóa.

Mục tiêu cụ thể

(a) Cung cấp một khuôn khổ chung về các giá trị, nguyên tắc và hành động để hướng dẫn các quốc gia xây dựng luật pháp, chính sách hoặc các công cụ khác liên quan đến AI, phù hợp với luật pháp quốc tế;

b) Hướng dẫn hành động của các cá nhân, nhóm, cộng đồng, tổ chức và công ty thuộc khu vực tư nhân nhằm bảo đảm việc đưa đạo đức vào tất cả các giai đoạn của vòng đời hệ thống AI;

(c) Bảo vệ, thúc đẩy và tôn trọng nhân quyền và các quyền tự do cơ bản, nhân phẩm và bình đẳng, bao gồm cả bình đẳng giới; bảo vệ lợi ích của các thế hệ hiện tại và tương lai; bảo tồn môi trường, đa dạng sinh học và các hệ sinh thái; và tôn trọng sự đa dạng văn hóa trong tất cả các giai đoạn của vòng đời hệ thống AI;

(d) Thúc đẩy đối thoại và xây dựng sự đồng thuận giữa nhiều bên liên quan, đa ngành về các vấn đề đạo đức liên quan đến hệ thống AI;

(e) Thúc đẩy khả năng tiếp cận công bằng với sự phát triển và kiến thức trong lĩnh vực AI và chia sẻ lợi ích, đặc biệt chú ý đến nhu cầu và đóng góp của quốc gia thu nhập trung bình.

2.2. Các giá trị và nguyên tắc

Các giá trị và nguyên tắc của Khuyến nghị phải được tất cả các tác nhân trong vòng đời của hệ thống AI tôn trọng ngay từ đầu và phù hợp, phải được thúc đẩy thông qua sửa đổi các luật, quy định và hướng dẫn kinh doanh mới hiện có hay đang xây dựng. Điều này phải tuân thủ luật pháp quốc tế, bao gồm Hiến chương Liên hợp quốc và các nghĩa vụ nhân quyền của các quốc gia thành viên, đồng thời phải phù hợp với các mục tiêu bền vững về xã hội, chính trị, môi trường, giáo dục, khoa học và kinh tế đã được quốc tế thống nhất, chẳng hạn như Mục tiêu phát triển bền vững của Liên hợp quốc (SDG).

Các giá trị đóng vai trò mạnh mẽ trong việc thúc đẩy các lý tưởng trong việc hình thành các biện pháp chính sách và quy phạm pháp luật. Mặc dù bản thân tất cả các giá trị và nguyên tắc được nêu dưới đây đều đáng được mong muốn, nhưng trong bất kỳ bối cảnh thực tế nào, có thể có những khó khăn trong thực hiện. Trong bất kỳ tình huống nào, việc đánh giá theo bối cảnh sẽ là cần thiết để quản lý những khó khăn tiềm ẩn, có tính đến nguyên tắc cân xứng và tuân thủ các quyền con người cũng như các quyền tự do cơ bản. Trong mọi trường hợp, mọi hạn chế có thể có đối với nhân quyền và các quyền tự do cơ bản đều phải có cơ sở pháp lý, hợp lý, cần thiết và tương xứng, đồng thời phù hợp với nghĩa vụ của các quốc gia theo luật pháp quốc tế. Để hướng dẫn các tình huống như vậy một cách thận trọng thường sẽ cần có sự tham gia của nhiều bên liên quan thích hợp, sử dụng đối thoại xã hội, cũng như cân nhắc về mặt đạo đức, thẩm định và đánh giá tác động.

Độ tin cậy và tính toàn vẹn của vòng đời của hệ thống AI là điều cần thiết để bảo đảm rằng công nghệ AI sẽ hoạt động vì lợi ích của nhân loại, cá nhân, xã hội, môi trường và hệ sinh thái, đồng thời thể hiện các giá trị và nguyên tắc được nêu trong Khuyến nghị này. Cần bảo đảm rằng hệ thống AI mang lại lợi ích cá nhân và lợi ích chung, đồng thời thực hiện các biện pháp thích hợp để giảm thiểu rủi ro. Một yêu cầu thiết yếu để có được độ tin cậy là trong suốt vòng đời của chúng, các hệ thống AI phải chịu sự giám sát kỹ lưỡng của các bên liên quan. Vì độ tin cậy là kết quả của việc vận hành các nguyên tắc trong Khuyến nghị này nên các hành động chính sách được đề xuất trong Khuyến nghị đều nhằm mục đích thúc đẩy độ tin cậy trong tất cả các giai đoạn của vòng đời hệ thống AI.

Giá trị

- Tôn trọng, bảo vệ và thúc đẩy nhân quyền, các quyền tự do cơ bản và phẩm giá con người: đây là cốt lõi trong hệ thống quyền con người và tự do cơ bản. Nó phản ánh sự cần thiết của việc bảo đảm rằng AI không gây tổn thương hoặc làm suy yếu nhân phẩm và quyền con người, và vai trò của các cá nhân và các tập thể khác nhau trong việc bảo đảm điều này.

- Môi trường và hệ sinh thái phát triển cân bằng và phồn thịnh: bảo vệ môi trường và thúc đẩy sự cân bằng và phồn thịnh của hệ sinh thái trong suốt quá trình vận hành của các hệ thống AI. Môi trường và hệ sinh thái là điều kiện cần để con người và các sinh vật khác có thể tận hưởng những lợi ích từ sự tiến bộ của AI; tuân thủ luật pháp quốc tế và quy định trong nước liên quan đến bảo vệ môi trường và hệ sinh thái, đồng thời kêu gọi giảm thiểu tác động môi trường của các hệ thống AI.

- Bảo đảm tính đa dạng và toàn diện: không hạn chế lựa chọn và quyền sử dụng AI của mọi người, và đồng thời cần có nỗ lực để giúp các cộng đồng có cơ sở hạ tầng công nghệ và pháp lý đầy đủ; bảo đảm sự tôn trọng, bảo vệ và thúc đẩy sự đa dạng trong suốt quá trình vận hành của các hệ thống AI, tuân thủ pháp luật quốc tế, bao gồm cả luật nhân quyền.

- Sống trong xã hội hòa bình, công bằng và gắn kết với nhau: nhấn mạnh vai trò của các bên liên quan đến AI trong việc xây dựng các xã hội hòa bình và công bằng, dựa trên một tương lai liên kết với lợi ích của tất cả, tuân thủ nhân quyền và tự do cơ bản; thúc đẩy hòa bình, tính bao dung và công bằng trong việc sử dụng AI, đồng thời bảo vệ tự do và an toàn của con người.

Nguyên tắc

- Phù hợp và không gây hại: quy trình AI chỉ nên đáp ứng mục tiêu chính đáng và phù hợp với bối cảnh; cần đánh giá rủi ro và áp dụng biện pháp ngăn ngừa khi có nguy cơ gây hại cho con người, xã hội, hoặc môi trường; việc sử dụng AI cần bảo đảm phhương pháp AI phù hợp và tương xứng với mục tiêu, không vi phạm quyền con người; trong các quyết định quan trọng, cần có sự giám sát của con người; AI không nên dùng để chấm điểm xã hội hoặc giám sát đại trà.

- An toàn và bảo mật: những rủi ro không mong muốn và các lỗ hổng bị tấn công (rủi ro bảo mật) cần được tránh và phải được xử lý, ngăn ngừa và loại bỏ trong suốt vòng đời của hệ thống AI để bảo đảm an toàn cho con người, môi trường và hệ sinh thái; AI an toàn và bảo mật sẽ được hỗ trợ bởi việc phát triển các khung truy cập dữ liệu và bảo vệ quyền riêng tư.

- Công bằng và không phân biệt đối xử: các bên liên quan đến AI nên thúc đẩy công bằng xã hội và bảo đảm không phân biệt đối xử theo luật quốc tế. Lợi ích của AI cần được tiếp cận bởi tất cả mọi người, đặc biệt là các nhóm yếu thế. Các quốc gia thành viên cần thúc đẩy truy cập AI bao trùm, tôn trọng đa ngôn ngữ và đa dạng văn hóa, và giải quyết khoảng cách số. Các quốc gia phát triển nên hỗ trợ quốc gia kém phát triển để chia sẻ lợi ích AI, góp phần tạo ra trật tự thế giới công bằng hơn. Các bên liên quan đến AI cần nỗ lực giảm thiểu phân biệt đối xử và thiên lệch trong suốt vòng đời AI để bảo đảm công bằng.

- Tính bền vững: phát triển xã hội bền vững phụ thuộc vào việc đạt được các mục tiêu liên quan đến con người, xã hội, văn hóa, kinh tế và môi trường. Công nghệ AI có thể hỗ trợ hoặc cản trở các mục tiêu này, tùy thuộc vào cách chúng được áp dụng ở các quốc gia khác nhau. Do đó, cần liên tục đánh giá tác động của AI trên các khía cạnh này, phù hợp với các Mục tiêu Phát triển Bền vững (SDGs) của Liên hợp quốc.

- Quyền riêng tư và bảo vệ dữ liệu: quyền riêng tư, cần thiết để bảo vệ phẩm giá và quyền tự chủ của con người, phải được tôn trọng và bảo vệ suốt vòng đời của hệ thống AI. Dữ liệu cho AI phải được thu thập, sử dụng, chia sẻ, lưu trữ và xóa theo luật pháp quốc tế và các giá trị, nguyên tắc đề ra. Cần thiết lập các khung bảo vệ dữ liệu và cơ chế quản trị thích hợp ở cấp quốc gia và quốc tế, bảo vệ bởi hệ thống tư pháp, và tuân theo các tiêu chuẩn quốc tế về bảo vệ dữ liệu cá nhân. Các hệ thống thuật toán cần đánh giá tác động về quyền riêng tư, bao gồm cân nhắc xã hội và đạo đức, và áp dụng thiết kế bảo vệ quyền riêng tư. Các bên liên quan đến AI phải chịu trách nhiệm về việc bảo vệ thông tin cá nhân suốt vòng đời của hệ thống AI.

- Sự giám sát và tính quyết định của con người: sự giám sát và quyết định của con người trong các quy trình, hệ thống hoặc công nghệ mà con người tham gia để bảo đảm rằng quyết định cuối cùng và hành động được thực hiện đều được kiểm soát và quyết định bởi con người, chứ không phải bởi máy móc hoặc AI. Điều này giúp bảo đảm rằng các quyết định và hành động được thực hiện theo cách có tính nhân văn và đạo đức, và bảo đảm tính minh bạch, trách nhiệm và đáng tin cậy trong quy trình.

- Tính minh bạch và khả năng giải thích được: Sự minh bạch và khả năng giải thích được của hệ thống AI là cần thiết để bảo đảm tôn trọng quyền con người và tự do, đồng thời giúp tăng cường sự minh bạch và tin cậy của hệ thống. Minh bạch giúp giảm thấp phân biệt đối xử, trong khi khả năng giải thích được giúp người dùng hiểu rõ hơn về quyết định của hệ thống AI. Điều này đồng nghĩa với việc người dùng cần có quyền được thông tin đầy đủ và giải thích khi quyết định của hệ thống AI ảnh hưởng đến họ.

- Trách nhiệm và trách nhiệm pháp lý: Các bên liên quan đến AI và các quốc gia thành viên cần tôn trọng, bảo vệ và thúc đẩy quyền con người và tự do cơ bản, cũng như bảo vệ môi trường và hệ sinh thái, tuân thủ trách nhiệm đạo đức và pháp lý tương ứng. Cần phát triển các cơ chế giám sát, đánh giá tác động, kiểm toán và nỗ lực cần thiết để bảo đảm trách nhiệm cho các hệ thống AI và ảnh hưởng của chúng suốt vòng đời của chúng.

- Nhận thức và hiểu biết: cần tăng cường nhận thức và hiểu biết công cộng về AI, giá trị của dữ liệu thông qua giáo dục mở, tham gia cộng đồng, kỹ năng số, đạo đức AI và giáo dục truyền thông, với sự hợp tác của chính phủ, tổ chức quốc tế, xã hội dân sự, trường đại học, cơ quan truyền thông và doanh nghiệp. Việc học về tác động của AI cần bao gồm quan điểm về quyền con người và tự do cơ bản, cũng như về môi trường và hệ sinh thái.

2.3. Các lĩnh vực hành động chính sách

Các hành động chính sách được mô tả trong các lĩnh vực chính sách sau đây sẽ vận hành các giá trị và nguyên tắc được nêu trong Khuyến nghị này. Hành động chính là để các quốc gia thành viên đưa ra các biện pháp hiệu quả, chẳng hạn như khung chính sách hoặc cơ chế, và để bảo đảm rằng các bên liên quan khác, chẳng hạn như các công ty thuộc khu vực tư nhân, các tổ chức nghiên cứu và học thuật cũng như xã hội dân sự tuân thủ chúng bằng cách khuyến khích tất cả các bên liên quan phát triển nhân quyền, pháp quyền, dân chủ, các công cụ đánh giá tác động đạo đức và thẩm định phù hợp với Nguyên tắc Hướng dẫn của Liên hợp quốc về Kinh doanh và Nhân quyền. Quá trình xây dựng các chính sách hoặc cơ chế như vậy cần có sự tham gia của tất cả các bên liên quan và cần tính đến hoàn cảnh cũng như ưu tiên của mỗi quốc gia thành viên. UNESCO có thể là đối tác và hỗ trợ các quốc gia thành viên trong việc phát triển cũng như giám sát và đánh giá các cơ chế chính sách.

UNESCO thừa nhận rằng các quốc gia thành viên sẽ ở các giai đoạn sẵn sàng khác nhau để thực hiện Khuyến nghị này, về các mặt khoa học, công nghệ, kinh tế, giáo dục,

pháp lý, quy định, cơ sở hạ tầng, xã hội, văn hóa và các khía cạnh khác. Cần lưu ý rằng "sẵn sàng" ở đây là một trạng thái động. Do đó, để có thể thực hiện hiệu quả Khuyến nghị này, UNESCO sẽ: (1) phát triển phương pháp đánh giá mức độ sẵn sàng để hỗ trợ các quốc gia thành viên quan tâm xác định tình trạng của họ tại những thời điểm cụ thể; và (2) bảo đảm hỗ trợ các quốc gia thành viên quan tâm trong việc phát triển phương pháp đánh giá tác động đạo đức (EIA) của công nghệ AI của UNESCO, chia sẻ các thực tiễn tốt nhất và hướng dẫn đánh giá.

Đánh giá tác động đạo đức

- Các quốc gia thành viên nên đưa ra các khung đánh giá tác động, như đánh giá tác động đạo đức, để xác định và đánh giá các lợi ích, mối quan ngại và rủi ro của các hệ thống AI, cũng như các biện pháp bảo đảm trước các rủi ro, giảm nhẹ và giám sát phù hợp. Đánh giá tác động này nên xác định các tác động đối với quyền con người và tự do cơ bản, đặc biệt là những quyền của những người bị bất lợi và dễ tổn thương, quyền lao động, môi trường và hệ sinh thái, cũng như các yếu tố đạo đức và xã hội, và tạo điều kiện cho sự tham gia của công dân phù hợp với các giá trị và nguyên tắc được đề ra trong Khuyến nghị này.

- Các quốc gia và doanh nghiệp cần phát triển cơ chế giám sát để bảo đảm rằng các hệ thống AI không vi phạm quyền con người. Họ cũng cần đánh giá tác động kinh tế xã hội của AI để bảo đảm không tăng khoảng cách giữa giàu và nghèo cũng như khoảng cách số. Các biện pháp như kiểm tra thuật toán và giám sát quá trình triển khai là cần thiết, và đánh giá đạo đức của các hệ thống AI cần phải được minh bạch và có sự tham gia của cộng đồng. Chính phủ cần thiết lập khung pháp lý để bảo đảm đánh giá tác động đạo đức của AI và thiết lập các cơ chế giám sát phù hợp.

Quản trị và kiểm soát các vấn đề đạo đức AI

- Các quốc gia thành viên cần bảo đảm rằng cơ chế quản trị AI là minh bạch và liên quan đến các bên; nên dự báo, bảo vệ, giám sát, thúc đẩy tuân thủ và khắc phục bất kỳ thiệt hại nào do hệ thống AI gây ra để bảo vệ nhân quyền và nguyên tắc pháp luật. Cơ chế khắc phục từ cả khu vực công và tư, kèm theo việc thúc đẩy khả năng kiểm tra và theo dõi của các hệ thống AI.

- Các quốc gia thành viên cần tăng cường khả năng tổ chức và hợp tác với các nhà nghiên cứu để ngăn chặn việc sử dụng AI với mục đích xấu.

- Khuyến khích phát triển chiến lược AI quốc gia với các hình thức quản trị mềm như cơ chế chứng nhận, bảo đảm rằng chúng không gây hại cho sáng tạo hoặc bất lợi cho các tổ chức nhỏ hơn. Việc giám sát định kỳ là quan trọng để duy trì tính toàn vẹn của hệ thống và tuân thủ nguyên tắc đạo đức suốt vòng đời của hệ thống AI.

- Các cơ quan công cần tự đánh giá minh bạch về các hệ thống AI hiện có và đề xuất, bao gồm việc đánh giá xem việc áp dụng AI có phù hợp không, cần phải có thêm đánh giá để xác định việc áp dụng đó có dẫn đến vi phạm hoặc lạm dụng các nghĩa vụ luật nhân quyền của các quốc gia thành viên không, và nếu có thì cần phải cấm sử dụng.

- Các quốc gia nên khuyến khích các tổ chức công, tư nhân và xã hội dân sự liên quan đến quản trị AI có một Quản trị viên đạo đức AI độc lập hoặc các cơ chế khác để giám sát đánh giá ảnh hưởng đạo đức và giám sát liên tục để hướng dẫn đạo đức của các hệ thống AI; khuyến khích phát triển hệ sinh thái số cho sự phát triển đạo đức AI ở cấp quốc gia, đồng thời đóng góp vào sự hợp tác quốc tế; thiết lập cơ chế để bảo đảm sự tham gia tích cực, đặc biệt là các quốc gia đang phát triển, trong các cuộc thảo luận quốc tế về quản trị AI.

- Việc soạn thảo pháp luật quốc gia mới về các hệ thống AI phải tuân thủ các nghĩa vụ luật nhân quyền của các quốc gia thành viên và thúc đẩy quyền con người và tự do cơ bản trong suốt vòng đời của hệ thống AI.

- Các quốc gia nên cung cấp cơ chế để giám sát tác động xã hội và kinh tế của các hệ thống AI được sử dụng cho các trường hợp nhạy cảm với quyền con người, thông qua các cơ quan giám sát độc lập và cơ quan công cộng có trách nhiệm; tăng cường năng lực của hệ thống tòa án để đưa ra quyết định liên quan đến các hệ thống AI theo nguyên tắc pháp luật và phù hợp với các tiêu chuẩn quốc tế, bao gồm việc sử dụng các hệ thống AI trong các phiên xử, đồng thời bảo đảm nguyên tắc giám sát con người được thực hiện.

Chính sách dữ liệu

- Các quốc gia thành viên cần phát triển chiến lược quản lý dữ liệu bảo đảm đánh giá liên tục chất lượng dữ liệu huấn luyện cho hệ thống AI, bảo mật dữ liệu, và các cơ chế phản hồi để học hỏi từ sai lầm và chia sẻ thực tiễn tốt nhất.

- Các biện pháp bảo vệ quyền riêng tư phải tuân thủ luật quốc tế. Cần khuyến khích tất cả các bên liên quan đến AI, trong đó có doanh nghiệp, tuân thủ các tiêu chuẩn quốc tế và thực hiện đánh giá tác động riêng tư và đạo đức.

- Các quốc gia cần bảo đảm cá nhân giữ quyền kiểm soát dữ liệu cá nhân với các biện pháp bảo vệ phù hợp, tính minh bạch, cơ chế trách nhiệm và khả năng truy cập, xóa dữ liệu. Cần có giám sát độc lập để bảo vệ dữ liệu cá nhân và thúc đẩy dòng chảy thông tin quốc tế.

- Chính sách dữ liệu nên bảo đảm an ninh cho dữ liệu cá nhân và nhạy cảm, đặc biệt là dữ liệu về tội phạm, sinh trắc học, di truyền và sức khỏe. Cần thúc đẩy dữ liệu mở và điều chỉnh chính sách để hỗ trợ chia sẻ dữ liệu an toàn và hợp pháp.

- Cần sử dụng các bộ dữ liệu chất lượng và đa dạng cho AI, đầu tư vào tạo các bộ dữ liệu tiêu chuẩn vàng và khuyến khích chuẩn hóa dữ liệu. Các quốc gia nên thúc đẩy hợp tác công-tư trong chia sẻ dữ liệu chất lượng trong không gian dữ liệu an toàn.

Chính sách phát triển và hợp tác quốc tế

- Các quốc gia và tập đoàn quốc tế nên ưu tiên đạo đức AI bằng cách thảo luận các vấn đề đạo đức liên quan đến AI trong các diễn đàn quốc tế, liên chính phủ và đa bên.

- AI trong các lĩnh vực phát triển như giáo dục, khoa học, văn hóa, y tế, nông nghiệp, môi trường và kinh tế cần tuân thủ các giá trị và nguyên tắc của Khuyến nghị

này.

- Các quốc gia nên hợp tác thông qua các tổ chức quốc tế để cung cấp nền tảng cho sự hợp tác quốc tế về AI, đặc biệt hỗ trợ các nước thu nhập thấp, kém phát triển, không có biển và đảo nhỏ bằng cách đóng góp chuyên môn, tài trợ, dữ liệu và cơ sở hạ tầng.

- Các quốc gia cần thúc đẩy hợp tác quốc tế về nghiên cứu và đổi mới AI, tạo điều kiện cho các trung tâm và mạng lưới nghiên cứu, khuyến khích sự tham gia và lãnh đạo của các nhà nghiên cứu từ các nước thu nhập thấp và kém phát triển.

- Nghiên cứu đạo đức AI nên được thúc đẩy thông qua sự tham gia của các tổ chức quốc tế và các tập đoàn, để bảo đảm sử dụng AI một cách đạo đức bởi cả các thực thể công và tư, và phát triển các giải pháp công nghệ phù hợp với các khung đạo đức cụ thể.

- Cần khuyến khích hợp tác và trao đổi công nghệ quốc tế để thu hẹp khoảng cách công nghệ, tôn trọng luật pháp quốc tế và thúc đẩy sự trao đổi giữa các quốc gia, khu vực công và tư, và giữa các nước phát triển và kém phát triển.

Chính sách môi trường và hệ sinh thái

- Các quốc gia và doanh nghiệp cần đánh giá và giảm thiểu tác động môi trường trực tiếp và gián tiếp của hệ thống AI, bao gồm lượng khí thải carbon, tiêu thụ năng lượng và tác động từ khai thác nguyên liệu. Tất cả các bên liên quan đến AI phải tuân thủ luật và chính sách môi trường.

- Các quốc gia nên tạo động lực phát triển và áp dụng các giải pháp AI dựa trên quyền và đạo đức nhằm tăng cường khả năng chống chịu thiên tai, bảo vệ và tái tạo môi trường, và duy trì hành tinh. Chính sách AI nên bao gồm cả sự tham gia của cộng đồng địa phương và bản địa, hỗ trợ kinh tế tuần hoàn và tiêu dùng bền vững.

Ví dụ về ứng dụng AI đối với chính sách môi trường và hệ sinh thái: Bảo vệ, giám sát và quản lý tài nguyên thiên nhiên; Dự đoán, phòng ngừa và giảm thiểu vấn đề liên quan đến khí hậu; Tăng cường hệ sinh thái thực phẩm bền vững; Thúc đẩy sử dụng năng lượng bền vững; Phát triển hạ tầng, mô hình kinh doanh và tài chính bền vững; Phát hiện và dự đoán ô nhiễm, hỗ trợ các biện pháp giảm thiểu ô nhiễm.

- Khi chọn phương pháp AI, cần ưu tiên các phương pháp tiết kiệm dữ liệu, năng lượng và tài nguyên. Phải có bằng chứng cho thấy AI sẽ có hiệu quả như dự định hoặc các biện pháp bảo vệ đi kèm sẽ hỗ trợ cho việc sử dụng AI. Nếu không thể, nên áp dụng nguyên tắc phòng ngừa và tránh sử dụng AI nếu có tác động tiêu cực không cân xứng đến môi trường.

Chính sách về giới

- Các quốc gia cần tối đa hóa tiềm năng của công nghệ số và AI để đạt được bình đẳng giới, bảo đảm quyền và an toàn của phụ nữ và trẻ em gái trong suốt vòng đời của hệ thống AI. Đánh giá tác động đạo đức phải bao gồm góc nhìn giới.

- Cần có quỹ công để tài trợ cho các chương trình đáp ứng giới, kế hoạch hành động giới trong chính sách số quốc gia và chính sách hỗ trợ phụ nữ trong kinh tế số. Đầu tư

vào các chương trình tăng cường cơ hội cho phụ nữ trong khoa học, công nghệ, kỹ thuật và toán học (STEM) và công nghệ thông tin và truyền thông (ICT), nâng cao năng lực việc làm và phát triển nghề nghiệp cho phụ nữ.

- AI cần được sử dụng để thu hẹp các khoảng cách giới hiện có, bao gồm khoảng cách lương, như trong các ngành nghề và vị trí quản lý, giáo dục, tiếp cận và sử dụng công nghệ số.

- Tránh đưa các định kiến và phân biệt giới vào hệ thống AI, thay vào đó cần xác định và sửa chữa những vấn đề này. Cần tránh tác động tiêu cực của sự chia rẽ công nghệ đối với bình đẳng giới và bạo lực như quấy rối và buôn người, cả trực tuyến và ngoại tuyến.

- Khuyến khích nữ doanh nhân và sự tham gia của phụ nữ trong mọi giai đoạn của vòng đời AI bằng cách cung cấp và thúc đẩy các ưu đãi kinh tế và chính sách hỗ trợ. Bảo đảm quỹ công và tư dành cho các chương trình bao trùm giới và môi trường làm việc không quấy rối.

- Thúc đẩy đa dạng giới trong nghiên cứu AI qua việc cung cấp các ưu đãi cho phụ nữ tham gia vào lĩnh vực này, chống lại định kiến giới và quấy rối trong cộng đồng nghiên cứu AI, và khuyến khích chia sẻ thực tiễn tốt nhất về tăng cường đa dạng giới. UNESCO có thể giúp tạo kho lưu trữ các thực tiễn tốt nhất để khuyến khích sự tham gia của phụ nữ và các nhóm ít được đại diện trong mọi giai đoạn của vòng đời hệ thống AI.

Chính sách văn hóa

- Các quốc gia nên áp dụng hệ thống AI vào bảo tồn, phát triển và quản lý di sản văn hóa, bao gồm ngôn ngữ và tri thức bản địa. Cần có chương trình giáo dục về AI trong các lĩnh vực này và bảo đảm sự tham gia của các tổ chức và công chúng.

- Các quốc gia cần đánh giá tác động văn hóa của AI, đặc biệt là ứng dụng xử lý ngôn ngữ tự nhiên, để tối ưu hóa lợi ích và giảm thiểu tác động tiêu cực như sự biến mất của ngôn ngữ và biến thể văn hóa.

- Thúc đẩy giáo dục AI và đào tạo kỹ thuật số cho nghệ sĩ và chuyên gia sáng tạo để đánh giá và sử dụng AI trong công việc của họ, bảo tồn di sản văn hóa, đa dạng và tự do nghệ thuật; khuyến khích các doanh nghiệp địa phương trong lĩnh vực văn hóa sử dụng công cụ AI.

- Hợp tác với công ty công nghệ và các bên liên quan để bảo đảm cung cấp và tiếp cận đa dạng các biểu đạt văn hóa, đồng thời tăng cường khả năng tìm thấy nội dung địa phương qua các đề xuất thuật toán.

- Thúc đẩy nghiên cứu về AI và sở hữu trí tuệ (IP), bảo vệ quyền IP cho các tác phẩm tạo ra bởi AI và đánh giá tác động của AI đối với quyền lợi của chủ sở hữu IP.

- Khuyến khích bảo tàng, thư viện và kho lưu trữ sử dụng AI để nâng cao và tiếp cận bộ sưu tập của họ.

Chính sách giáo dục và nghiên cứu

- Các quốc gia cần hợp tác với tổ chức quốc tế, cơ sở giáo dục, và tổ chức phi chính phủ để cung cấp giáo dục AI cho công chúng nhằm giảm thiểu khoảng cách số và bất bình đẳng truy cập số.

- Thúc đẩy kỹ năng cơ bản như đọc viết, toán học, lập trình, kỹ năng số và kỹ năng tư duy phê phán, sáng tạo, làm việc nhóm, giao tiếp, kỹ năng xã hội và đạo đức AI, đặc biệt ở những khu vực thiếu hụt.

- Tăng cường nhận thức về phát triển AI và tác động của nó đến quyền con người, bao gồm quyền trẻ em.

- Khuyến khích nghiên cứu về sử dụng AI có trách nhiệm trong giảng dạy, đào tạo giáo viên và học trực tuyến, đánh giá chất lượng giáo dục và tác động của AI đối với học sinh và giáo viên, bảo đảm AI hỗ trợ mà không làm giảm khả năng nhận thức hay lạm dụng dữ liệu cá nhân.

- Thúc đẩy sự tham gia của phụ nữ, người khuyết tật và các nhóm yếu thế trong chương trình giáo dục AI, chia sẻ thực tiễn tốt nhất với các quốc gia khác.

- Phát triển chương trình đạo đức AI ở mọi cấp học, kết hợp giáo dục kỹ thuật AI với khía cạnh nhân văn, đạo đức và xã hội. Cung cấp khóa học trực tuyến và tài liệu giáo dục đạo đức AI bằng nhiều ngôn ngữ, bảo đ đảm ịnh dạng phù hợp cho người khuyết tật.

- Hỗ trợ và đầu tư vào nghiên cứu AI, đặc biệt là đạo đức AI, và khuyến khích sự hợp tác giữa các nhà nghiên cứu và công ty phát triển AI có đạo đức.

- Bảo đảm rằng các nhà nghiên cứu AI được đào tạo về đạo đức nghiên cứu và tích hợp các yếu tố đạo đức vào thiết kế, sản phẩm và công bố của họ, đặc biệt trong phân tích dữ liệu.

- Khuyến khích công ty tư nhân chia sẻ dữ liệu cho nghiên cứu, tuân thủ tiêu chuẩn bảo vệ quyền riêng tư và dữ liệu.

- Bảo đảm phát triển AI dựa trên nghiên cứu khoa học độc lập, thúc đẩy nghiên cứu liên ngành và nhận thức về lợi ích, giới hạn và rủi ro của công nghệ AI, hỗ trợ cộng đồng khoa học đóng góp vào chính sách và nhận thức về AI.

Chính sách truyền thông và thông tin

- Các quốc gia nên sử dụng hệ thống AI để cải thiện truy cập thông tin và tri thức, hỗ trợ nhà nghiên cứu, học giả, nhà báo, công chúng và nhà phát triển, tăng cường tự do ngôn luận và tự do học thuật.

- Bảo đảm các hệ thống AI tôn trọng và thúc đẩy tự do ngôn luận, truy cập thông tin, và minh bạch trong việc tạo nội dung, kiểm duyệt và quản lý nội dung trực tuyến. Cung cấp khung pháp lý cho người dùng tiếp cận nhiều quan điểm và có cơ chế khiếu nại.

- Đầu tư và thúc đẩy kỹ năng số, truyền thông và thông tin để tăng cường tư duy phản biện, hiểu biết về AI, giảm thiểu thông tin sai lệch và ngôn từ kích động thù địch.

- Tạo môi trường hỗ trợ truyền thông báo cáo về lợi ích và tác hại của AI, khuyến

khích sử dụng AI có đạo đức trong các hoạt động truyền thông.

Chính sách kinh tế và lao động

- Các quốc gia nên đánh giá và giải quyết tác động của hệ thống AI lên thị trường lao động và yêu cầu giáo dục. Điều này bao gồm việc giảng dạy các kỹ năng cốt lõi và liên ngành, như học cách giao tiếp, tư duy phản biện, làm việc nhóm và đồng cảm, bên cạnh các kỹ năng chuyên môn và kỹ thuật.

- Hỗ trợ các thỏa thuận hợp tác giữa chính phủ, khu vực hàn lâm, các cơ sở giáo dục nghề nghiệp, ngành công nghiệp, tổ chức lao động và xã hội dân sự để điều chỉnh chương trình đào tạo theo nhu cầu thị trường và tương lai công việc. Thúc đẩy phương pháp dạy và học dựa trên dự án, cho phép hợp tác giữa các tổ chức công, công ty tư nhân, đại học và trung tâm nghiên cứu.

- Làm việc với công ty tư nhân, tổ chức xã hội dân sự và các bên liên quan để bảo đảm quá trình chuyển đổi công bằng cho nhân viên có nguy cơ thất nghiệp, bao gồm các chương trình nâng cao kỹ năng và đào tạo lại, cơ chế duy trì nhân viên và các chương trình an sinh xã hội.

- Khuyến khích nghiên cứu tác động của AI lên môi trường lao động địa phương để dự đoán xu hướng và thách thức, đồng thời đề xuất các phương pháp tốt nhất cho việc đào tạo lại và bố trí lại nhân lực.

- Bảo đảm thị trường cạnh tranh và bảo vệ người tiêu dùng, ngăn chặn lạm dụng vị thế thống trị của các công ty, đặc biệt ở các nước kém phát triển. AI phải tuân thủ các tiêu chuẩn đạo đức khi xuất khẩu hoặc phát triển ở các quốc gia chưa có tiêu chuẩn này.

Chính sách y tế và phúc lợi xã hội

- Các quốc gia nên sử dụng hệ thống AI hiệu quả để cải thiện sức khỏe con người và bảo vệ quyền sống, bao gồm giảm nhẹ các đợt bùng phát dịch bệnh, đồng thời xây dựng và duy trì sự đoàn kết quốc tế để đối phó với các rủi ro và không chắc chắn về sức khỏe toàn cầu, bảo đảm rằng triển khai hệ thống AI trong chăm sóc sức khỏe phù hợp với luật pháp quốc tế và các nghĩa vụ về nhân quyền.

- Quan tâm đến việc quy định các giải pháp dự báo, phát hiện và điều trị trong các ứng dụng AI bằng cách: (a) bảo đảm giám sát để giảm thiểu sự thiên vị; (b) bảo đảm bác sĩ, bệnh nhân, người chăm sóc hoặc người dùng dịch vụ được bao gồm trong tất cả các bước liên quan đến việc phát triển các thuật toán; (c) chú ý đặc biệt đến quyền riêng tư cần thiết cho việc được theo dõi y tế và bảo đảm rằng tất cả các yêu cầu về bảo vệ dữ liệu quốc gia và quốc tế liên quan được đáp ứng; (d) bảo đảm các cơ chế hiệu quả để những người có dữ liệu cá nhân có ý thức và đồng ý cho sử dụng và phân tích dữ liệu của họ, không ngăn cản quyền truy cập vào chăm sóc sức khỏe; (e) bảo đảm sự quan tâm con người và quyết định cuối cùng về chẩn đoán và điều trị luôn được thực hiện bởi con người trong khi công nhận rằng hệ thống AI cũng có thể hỗ trợ trong công việc; (f) bảo đảm rằng hệ thống AI cần được xem xét bởi một ủy ban nghiên cứu đạo đức trước khi đưa vào sử dụng lâm sàng.

Tóm lại

Đạo đức AI là một vấn đề phức tạp và có quy mô toàn cầu, thu hút sự quan tâm của nhiều quốc gia và tổ chức quốc tế. Dựa trên kinh nghiệm từ nhiều quốc gia và khuyến nghị của Liên hợp quốc, Việt Nam đang tiến hành xây dựng các quy định nhằm phát triển AI một cách có đạo đức và trách nhiệm.

Theo Liên hợp quốc, việc phát triển AI có đạo đức đòi hỏi sự quản lý cẩn trọng từ việc xác định mô hình AI, thu thập dữ liệu, đến hoàn thiện hệ thống và đưa vào ứng dụng. Để bảo đảm AI phát triển một cách có trách nhiệm, cần sự phối hợp của nhiều bên, đặc biệt là các cơ quan quản lý, các tổ chức, cá nhân nghiên cứu và phát triển AI, cùng sự tham gia của cộng đồng quốc tế, khu vực và quốc gia.

Mô hình AI cần tuân thủ các thiết kế và nhiệm vụ đã được thiết lập từ đầu, bảo đảm không có hành động phá hoại hay gây tổn hại cho con người. Khác với các công nghệ trước đây, AI có khả năng tự tạo ra hướng đi mới, vượt ngoài sự kiểm soát của nhà phát triển. Do vậy, các vấn đề bình đẳng, công bằng, minh bạch, an toàn, bảo mật, tôn trọng các quyền của con người, luật pháp quốc tế… cần được chú ý khi xây dựng mô hình AI.

Việc xây dựng và áp dụng các quy tắc đạo đức AI là một nhiệm vụ quan trọng, cần bảo đảm rằng AI phát triển không chỉ hiệu quả mà còn đáp ứng các tiêu chuẩn đạo đức và trách nhiệm xã hội. AI cần hoạt động vì lợi ích của nhân loại, cá nhân, xã hội, môi trường và hệ sinh thái, đồng thời phòng ngừa các rủi ro, tác hại. Do sự phức tạp của các vấn đề pháp lý và đạo đức xung quanh AI nên sự hợp tác của nhiều bên liên quan ở các cấp độ và lĩnh vực khác nhau là cần thiết trong quá trình phát triển và ứng dụng AI.