IPS Viện Nghiên cứu Chính sách và Phát triển Truyền thông

CHÍNH SÁCH PHÁT TRIỂN VÀ KIỂM SOÁT RỦI RO ĐỐI VỚI AI

THÁNG 10 NĂM 2024

A: 18th Floor, VTC Online Building, 18 Tam Trinh, Hanoi
T: (+84) 373 643 601 | E: contact@ips.org.vn | W: https://ips.org.vn/

Giới thiệu về Viện Nghiên cứu Chính sách và Phát triển Truyền thông (IPS)

Viện Nghiên cứu Chính sách và Phát triển Truyền thông (IPS) là tổ chức nghiên cứu chính sách độc lập thuộc Hội Truyền thông số Việt Nam. IPS cung cấp các phân tích và giải pháp chính sách nhằm tối ưu tiềm năng của công nghệ số, góp phần thúc đẩy thịnh vượng kinh tế và củng cố an ninh cho Việt Nam cũng như khu vực Đông Nam Á. Lĩnh vực nghiên cứu của IPS gồm kinh tế số, chính phủ số và xã hội số.

Về kinh tế số, IPS tập trung vào nghiên cứu (1) chính sách về hạ tầng cho kinh tế số như cáp quang biển, vệ tinh, và trung tâm dữ liệu và (2) chính sách điều tiết thị trường dịch vụ số, nhằm tăng cường giao lưu thương mại, văn hóa, kinh tế giữa Việt Nam với các nước trên thế giới. Về xã hội số, IPS đi sâu nghiên cứu thực tiễn bảo vệ quyền trên môi trường số, khuyến khích sự tham gia của công dân vào quản trị công. Về chính phủ số, IPS tiên phong trong nghiên cứu quản trị dữ liệu, thúc đẩy dữ liệu mở và nâng cao hiệu quả dịch vụ công, tạo nên những giải pháp đột phá nhằm mang lại lợi ích cho mọi người dân.

A: 18th Floor, VTC Online Building, 18 Tam Trinh, Hanoi
T: (+84) 373 643 601 | E: contact@ips.org.vn | W: https://ips.org.vn/

Tóm tắt

Tài liệu này thảo luận về các cách tiếp cận chính sách nhằm Quản lý rủi ro do công nghệ Trí tuệ nhân tạo tạo ra và khuyến nghị cụ thể cho tiếp cận pháp lý về quản lý rủi ro AI trong Luật Công nghiệp công nghệ số. Các điểm chính được phân tích trong báo cáo gồm:

- Công nghệ Trí tuệ nhân tạo (AI) có tiềm năng mang lại lợi ích lớn lao cho phát triển kinh tế - xã hội ở Việt Nam, trong đó nổi bật nhất là tiềm năng nâng cao năng suất lao động; cải thiện năng suất, hiệu quả các ngành kinh tế hiện hữu.

- Từ góc độ rủi ro, công nghệ này cũng có thể tạo ra các tác động tiêu cực cả trên 2 bình diện cá nhân/công dân và xã hội. Cụ thể, rủi ro cho cá nhân/công dân: về (i) cơ hội việc làm (do AI thay thế một phần/hoàn toàn lao động con người trong một số loại hình công việc cụ thể); (ii) tác động đến an toàn cá nhân và quyền công dân, cụ thể gồm tác động đến (a) phẩm giá con người (mất danh tính; mất tương tác cá nhân); (b) tự do (bị thao túng và kiểm soát hành vi, bị giám sát theo dõi, hạn chế không gian riêng tư); (c) gia tăng bất bình đẳng (thiên kiến và bị phân biệt đối xử). Ở cấp độ xã hội, tác động tiêu cực đáng lo ngại nhất là (a) sự phân mảnh xã hội, giảm niềm tin, tính gắn kết cộng đồng và (b) bỏ qua các nhóm dễ bị tổn thương.

- Lợi ích lẫn rủi ro lớn lao mà AI tạo ra đang đòi hỏi các Chính phủ phải có các chính sách can thiệp vào sự phát triển của bản thân công nghệ AI lẫn thị trường (ứng dụng) công nghệ.

- Đối với Việt Nam, lợi ích của AI ở Việt Nam trên thực tế vẫn ở mức tiềm năng. Để biến tiềm năng đó thành hiện thực, cần có những can thiệp chính sách mang tính khuyến khích trong giai đoạn 3-5 sắp tới. Hơn thế nữa, do công nghệ AI vẫn đang tiếp tục phát triển trên bình diện toàn cầu, đưa ra quy định pháp lý 'cứng' dễ khiến quy định bị 'lạc hậu', cản trở phát triển công nghệ.

- Do đó, Việt Nam nên chọn cách xây dựng quy định pháp lý theo hướng 'nới lỏng', khuyến khích tổ chức, doanh nghiệp, cá nhân thực hiện cơ chế tuân thủ tự nguyện qua xây dựng và thực hiện các bộ quy tắc ứng xử hoặc tuân thủ các bộ tiêu chuẩn công nghệ do ngành công nghệ đặt ra.

- Cụ thể, nội dung về AI trong dự thảo Luật Công nghiệp công nghệ số nên đưa ra các nguyên tắc về phát triển, ứng dụng AI an toàn; quy định trách nhiệm về việc xây dựng, thực hiện các bộ quy tắc ứng xử về đạo đức và an toàn AI. Luật chưa nên đưa ra các nghĩa vụ pháp lý mang tính bắt buộc phải thực thi cho các chủ thể.

A: 18th Floor, VTC Online Building, 18 Tam Trinh, Hanoi
T: (+84) 373 643 601 | E: contact@ips.org.vn | W: https://ips.org.vn/

I. Nhận diện các rủi ro pháp lý của công nghệ Trí tuệ nhân tạo

Việc tích hợp AI vào các lĩnh vực khác nhau có thể thách thức việc bảo vệ và thúc đẩy các giá trị cốt lõi của quyền con người: phẩm giá, tự do, bình đẳng và đoàn kết/gắn kết xã hội.

(a) Tác động của AI lên phẩm giá

Mất danh tính: Hệ thống AI như chấm điểm công dân có thể khiến con người trở nên vô danh khi chỉ được coi như các điểm dữ liệu, biến số thống kê hoặc như máy móc. Các quy trình ra quyết định tự động thiếu minh bạch có thể khiến con người bị đối xử một cách phi cá nhân, bỏ qua sự phức tạp trong danh tính và trải nghiệm của họ. Quan điểm giản đơn này có thể làm mất đi giá trị của con người, xói mòn sự tôn trọng đối với giá trị nội tại của con người.

Mất sự tương tác cá nhân: Trong các lĩnh vực như y tế hoặc dịch vụ khách hàng, AI có thể thay thế sự tương tác giữa con người, dẫn đến cảm giác không được tôn trọng và chăm sóc đúng mức. Sự thay đổi này có thể khiến mọi người cảm thấy như hoàn cảnh riêng của họ bị bỏ qua.

(b) Tác động của AI lên tự do

Thao túng và kiểm soát hành vi: AI có thể được sử dụng để ảnh hưởng hoặc thao túng hành vi, đặc biệt thông qua quảng cáo có mục tiêu hoặc nội dung cá nhân hóa có thể hình thành ý kiến và quyết định của con người, xâm phạm quyền tự chủ của cá nhân.

Giám sát, theo dõi: Các hệ thống giám sát dựa trên AI có thể xâm phạm quyền tự do cá nhân bằng cách theo dõi và phân tích hành động của các cá nhân mà không có sự đồng ý của họ. Điều này có thể tạo ra hiệu ứng sợ hãi, khiến mọi người thay đổi hành vi vì sợ bị theo dõi.

(c) Tác động của AI lên bình đẳng

Thiên kiến và phân biệt đối xử: Các hệ thống AI được đào tạo trên dữ liệu thiên lệch có thể duy trì hoặc thậm chí làm trầm trọng thêm các bất bình đẳng hiện có. Ví dụ, các công nghệ nhận diện khuôn mặt đã được chứng minh là có tỷ lệ lỗi cao hơn đối với một số nhóm nhân khẩu học nhất định. Các thuật toán AI được sử dụng trong quy trình tuyển dụng hoặc tư pháp hình sự có thể củng cố những thiên kiến về chủng tộc, giới tính, kinh tế, xã hội có trong dữ liệu huấn luyện.

Hệ thống quyết định tự động: AI ngày càng được sử dụng trong các quy trình ra quyết định trong các lĩnh vực như chấm điểm tín dụng, thực thi pháp luật, và dịch vụ xã hội. Khi các hệ thống này đưa ra các quyết định ảnh hưởng đến cuộc sống của cá nhân, chẳng hạn như từ chối cho vay hoặc xác định nghi phạm, chúng có thể làm như vậy theo cách bất công hoặc phân biệt đối xử, đặc biệt nếu không được điều chỉnh hoặc kiểm soát đúng cách.

Tiếp cận lợi ích của AI: AI có thể làm sâu sắc thêm khoảng cách số, với các cộng đồng bị thiệt thòi có ít cơ hội tiếp cận lợi ích của các công nghệ AI. AI có thể làm trầm trọng thêm các bất bình

đẳng hiện có bằng cách mang lại lợi ích cho những người có thể tiếp cận công nghệ trong khi gạt ra ngoài lề những người yếu thế.

(d) Tác động của AI lên gắn kết xã hội

Phân mảnh xã hội: AI, đặc biệt là trong các nền tảng truyền thông xã hội và thông tin, có thể dẫn đến phân mảnh xã hội bằng cách thúc đẩy nội dung củng cố các niềm tin hiện có và chia rẽ các cộng đồng.

Bỏ qua các nhóm dễ bị tổn thương: Các hệ thống AI có thể ưu tiên hiệu quả và lợi nhuận, bỏ qua nhu cầu của các nhóm dễ bị tổn thương. Ví dụ, các hệ thống tự động trong các dịch vụ xã hội có thể không nhạy cảm với sự phức tạp của từng trường hợp cụ thể, dẫn đến hỗ trợ không đầy đủ.

1.1. Rủi ro của AI đối với các quyền con người cụ thể

Các quyền con người cụ thể chịu tác động tiêu cực từ AI như quyền về đời tư, quyền biểu đạt, các quyền về giáo dục, y tế, quyền bình đẳng và không bị phân biệt đối xử, v.v… Nổi bật nhất và nghiêm trọng nhất là việc sử dụng AI một cách sai lệch có thể làm gia tăng sự phân biệt đối xử, vi phạm các quyền riêng tư, dữ liệu cá nhân, an toàn cá nhân.

Vi phạm quyền riêng tư: Các điều khoản về quyền riêng tư trong các văn kiện quốc tế về quyền con người mà công nghệ AI dễ vi phạm[1]: **Điều 12 của Tuyên ngôn quốc tế về quyền con người (UDHR):** không ai phải chịu sự can thiệp tùy ý vào cuộc sống riêng tư, gia đình, nơi ở hoặc thư tín hay bị xúc phạm danh dự hoặc uy tín cá nhân; **Điều 17 của Công ước quốc tế về quyền dân sự và chính trị (ICCPR):** Bảo vệ chống lại sự can thiệp tùy tiện vào quyền riêng tư, gia đình, nhà ở hoặc thư tín; **Điều 8 của Hiến chương EU về các quyền cơ bản:** Bảo đảm quyền bảo vệ dữ liệu, yêu cầu xử lý dữ liệu cá nhân một cách công bằng dựa trên sự đồng ý hoặc cơ sở hợp pháp.
Các hành động vi phạm quyền riêng tư như: Việc triển khai AI trong các hệ thống giám sát, chẳng hạn như công nghệ nhận diện khuôn mặt, có thể theo dõi cá nhân mà không có sự đồng ý của họ; các hệ thống AI thu thập dữ liệu của cá nhân mà không có sự đồng ý rõ ràng hoặc hiểu biết về cách dữ liệu của họ sẽ được sử dụng. Các chủ thể cả Nhà nước và tư nhân ở Việt Nam đang tăng cường giám sát bằng các công nghệ AI. Điều này có thể cải thiện trật tự, an toàn xã hội, nhưng cũng đặt ra lo ngại về vi phạm quyền riêng tư và rủi ro lạm dụng quyền lực của các cơ quan, tổ chức.

Tác động đối với quyền có việc làm (Điều 23 UDHR): Việc tích hợp các hệ thống hỗ trợ quyết định, robot trợ giúp và phương tiện tự lái có tiềm năng khiến nhiều người lao động mất việc, củng cố định kiến trên thị trường lao động, ảnh hưởng đến quyền được làm việc, dẫn đến bất ổn kinh tế và bất bình đẳng xã hội.

[1] D. Majumdar & H.K. Chattopadhyay, AI and Human Rights: From Business and Policy Perspectives.

Ảnh hưởng đến tự do biểu đạt (Điều 19 UDHR): Các thuật toán AI được các nền tảng truyền thông xã hội hoặc cơ quan công quyền sử dụng có thể ảnh hưởng và kiểm soát thông tin, dẫn đến việc kiểm duyệt và thao túng ý kiến công chúng; có thể ảnh hưởng quyền tự do biểu đạt và tiếp cận thông tin.

Hạn chế quyền tự do cá nhân: AI được sử dụng để theo dõi con người trên diện rộng, bằng các phần mềm nhận diện ngày càng tinh vi như phần mềm nhận diện khuôn mặt, nhận diện dáng đi…. nhằm phát hiện, đàn áp các nhóm và cá nhân nhất định, chấm điểm công dân[2]. Thực tế này đặt công dân vào tình trạng luôn cảm thấy bị theo dõi, không còn không gian riêng tư, tự do; cản trở họ thực hiện các quyền hợp pháp như tự do hiệp hội, hội họp, nhất là các nhóm dễ tổn thương. Hoặc là việc sử dụng AI ngày càng phổ biến trong hệ thống tư pháp hình sự ở các nước làm tăng nguy cơ xâm phạm các quyền tự do cá nhân. Ví dụ việc sử dụng phần mềm đánh giá rủi ro về việc tái phạm được sử dụng ở Mỹ để đưa ra quyết định về việc tạm giam hay cho tại ngoại, thậm chí xác định hình phạt với một người.

1.2. Rủi ro vi phạm quyền con người theo vòng đời của AI

Nguy cơ, rủi ro đối với quyền con người có thể xảy ra trong tất cả các khâu thuộc vòng đời của AI, từ việc thiết kế, sử dụng dữ liệu huấn luyện AI, cho đến người dùng cuối. Bảng dưới đây nêu các ví dụ về những rủi ro tiềm ẩn trực tiếp và gián tiếp như vậy[3].

Các giai đoạn vòng đời AI	Hành động có nguy cơ, rủi ro đối với quyền con người	Tác động tiêu cực đối với quyền con người
Thiết kế	Các nhà vận hành hệ thống và nhà thiết kế AI lập kế hoạch hoặc thiết kế hệ thống mà không tính đến những tác hại do các chế độ lỗi tiềm ẩn, và/hoặc do sử dụng vượt ra khỏi phạm vi dự kiến ban đầu. Các nhà vận hành và nhà thiết kế cố ý thiết kế một hệ thống, hoặc lựa chọn dữ liệu và thuật toán vi phạm quyền con người ngay từ khâu thiết kế, ví dụ như cho phép theo dõi tùy tiện và bất hợp pháp.	Các mô hình và ứng dụng AI có thể được sử dụng theo những cách dẫn đến phân biệt đối xử, không an toàn hoặc có hại, vượt ngoài ý định của nhà phát triển. Ví dụ, một công cụ xử lý hình ảnh không xử lý đúng các tông màu da sẫm có thể tạo ra sự thiên vị; hoặc một hệ thống được thiết kế để giám sát chuyển động của xe hoặc người có thể bị lạm dụng để xâm phạm quyền riêng tư hoặc quấy rối. Các tác động có thể xảy ra như: giám sát và theo dõi trái phép và/hoặc không chính xác (bao gồm quản lý bằng thuật toán không phù hợp và giám sát nơi làm việc do AI hỗ trợ); bắt giữ và giam giữ sai; tổn hại do tạo tài liệu lạm

[2] Nguyễn Văn Quân, Bùi Phú Châu, Ứng dụng trí tuệ nhân tạo và tác động tới quyền con người: Góc nhìn từ mô hình đánh giá tín nhiệm xã hội ở Trung Quốc, trong sách: NXB, 2020.
[3] Risk Management Profile for AI and Human Rights - United States Department of State

		dụng tình dục trẻ em hoặc hình ảnh riêng tư không có sự đồng ý. Có thể xảy ra việc hạn chế quyền tự do ngôn luận, tự do hội họp và lập hội một cách hòa bình.
Thu thập, xử lý, sử dụng dữ liệu huấn luyện	Dữ liệu được thu thập hoặc lấy từ các nguồn khác nhau để huấn luyện các mô hình AI, tái sử dụng cho các ứng dụng khác, hoặc bán mà người dùng không hề biết. Ví dụ, các cơ quan hoặc công ty bán hoặc chuyển giao dữ liệu sinh trắc học cho các quốc gia khác hoặc các công ty tư nhân. Độ chính xác của các tập dữ liệu dùng để huấn luyện các mô hình AI không được xác minh đầy đủ trước khi sử dụng. Các tập dữ liệu không đại diện đủ cho chủng tộc, giới tính, các đặc điểm cá nhân khác hoặc các khía cạnh văn hóa (bao gồm ngôn ngữ). Các tập dữ liệu có thể dựa trên nội dung hoặc các sự kiện thiên kiến trong thế giới thực, có khả năng củng cố sự bất bình đẳng hoặc những định kiến có hại.	Các mô hình AI có thể có tỷ lệ sai sót cao hơn hoặc không mang lại lợi ích cho những cá nhân có các đặc điểm không được đại diện đầy đủ trong dữ liệu huấn luyện, hoặc từ dữ liệu không chính xác. Nếu không được giải quyết một cách thích hợp, các đầu ra sai lệch của mô hình có thể dẫn đến các hậu quả đối với quyền con người như bắt giữ trái phép, từ chối phúc lợi xã hội, từ chối tín dụng, hoặc các hậu quả có hại khác. Trong lĩnh vực tư pháp hình sự, dữ liệu có thể được sử dụng để huấn luyện các mô hình AI cho phép các suy luận dự đoán, củng cố các khuôn mẫu phân biệt đối xử đã tồn tại, trong đó có việc phân biệt chủng tộc và sắc tộc. Các hệ thống có thể tạo ra kết quả sai (nhận diện sai) đối với cá nhân, dẫn đến bắt giữ oan. Các định kiến có hại và thiên lệch trong các tập dữ liệu có thể dẫn đến duy trì các định kiến hiện hành dựa trên chủng tộc, màu da, giới tính, ngôn ngữ, tôn giáo hoặc niềm tin, quan điểm chính trị, nguồn gốc quốc gia hoặc xã hội, tài sản, hoặc xuất thân, hoặc các đặc điểm khác, làm trầm trọng thêm sự phân biệt đối xử và bất bình đẳng kinh tế-xã hội hiện hành. Các tác động từ dữ liệu không chính xác, không đại diện hoặc có hại có thể gây ra hiệu ứng tiêu cực đối với cá nhân hoặc nhóm, những người không tin tưởng vào độ chính xác và hiệu quả của các hệ thống AI, hoặc lo sợ bị theo dõi, giám sát vì bày tỏ quan điểm ở nơi công cộng.
Xây dựng, kiểm tra, phê duyệt mô hình	Các nhà thiết kế hệ thống triển khai các biện pháp bảo vệ kỹ thuật không đủ để ngăn chặn rò	Các công cụ AI có thể suy luận thông tin nhận dạng cá nhân vi phạm quyền riêng tư, bao gồm các thuộc tính nhạy cảm như vị trí, giới tính,

	rỉ dữ liệu, tiết lộ trái phép, hoặc hủy ẩn danh thông tin nhận dạng cá nhân, hoặc các dữ liệu nhạy cảm khác như dữ liệu sinh trắc học, sức khỏe, chỗ ở. Các nhà phát triển hoặc các tác nhân khác không tiến hành thử nghiệm và đánh giá để phát hiện các kết quả đầu ra không chính xác, bao gồm cả các phản hồi thiên lệch hoặc bịa đặt. Tương tác giữa các mô hình AI và con người được thiết kế theo cách mà các câu trả lời của mô hình được cung cấp cho người dùng mà không kiểm tra đầy đủ, bao gồm việc không cung cấp và đánh giá các lời giải thích hoặc lý do cho các đầu ra của hệ thống "hộp đen".	tuổi tác, khuynh hướng tình dục và niềm tin chính trị. Người dùng có thể bị tái nhận dạng, có thể bằng cách kết hợp dữ liệu ẩn danh với các điểm dữ liệu khác và có thể bị theo dõi qua các vị trí vật lý và không gian trực tuyến. Người dùng có thể trích xuất dữ liệu huấn luyện từ các mô hình cho phép tái dựng thông tin định dạng cá nhân mà không có sự đồng ý của họ. Các mô hình AI có thể không thực hiện được chức năng dự kiến, gây hại cho những người mà việc tận hưởng quyền con người của họ phụ thuộc vào chức năng đó (ví dụ: dịch tài liệu liên quan đến việc xin tị nạn, hoặc đưa ra các quyết định có ảnh hưởng quan trọng đến cuộc sống). Các dịch vụ hoặc tài nguyên có thể bị từ chối hoặc thu hồi dựa trên các quyết định không có căn cứ hoặc dữ liệu không chính xác. Những người bị ảnh hưởng bởi các đầu ra hoặc quyết định của AI có thể gặp khó khăn trong việc xác định khi nào và cách thức họ bị tổn hại, dẫn đến giảm trách nhiệm giải trình và khả năng giảm thiểu hoặc khắc phục các tổn hại.
Triển khai, sử dụng	Các doanh nghiệp, cơ quan công bố hoặc triển khai một mô hình AI mới với thiếu sót nêu trên chưa được giải quyết, hoặc họ đưa AI vào ứng dụng mà không có hướng dẫn và biện pháp bảo vệ để đảm bảo sử dụng có trách nhiệm. Người dùng cuối vượt qua các biện pháp bảo vệ để sử dụng hệ thống AI vi phạm hoặc lạm dụng quyền con người.	Những rủi ro được nêu trên có thể trở thành hiện thực. Ví dụ như các mô hình AI có thể bị lạm dụng để tạo ra hình ảnh thân mật không có sự đồng thuận, hoặc tài liệu lạm dụng tình dục trẻ em, có thể được sử dụng để kiếm lợi, đe dọa nạn nhân, hoặc để tiến hành giám sát trái phép hoặc kiểm duyệt. Thông tin sai lệch được hỗ trợ bởi AI có thể được sử dụng để quấy rối và đe dọa nhà báo, đối thủ chính trị, hoặc những người bảo vệ nhân quyền, buộc họ phải tự kiểm duyệt. Điều này có thể làm suy yếu khả năng thực thi quyền tự do tìm kiếm và nhận thông tin cần thiết để hình thành quan điểm, cũng như quyền tự do hội họp và lập hội một cách hòa bình.

Vận hành, giám sát	Các cơ quan, doanh nghiệp không cung cấp cách thức cho mọi người báo cáo và biện pháp khắc phục đối với các lạm dụng, sai sót, hoặc sự cố liên quan đến hệ thống AI.	Các hệ thống AI có thể tiếp tục làm trầm trọng hơn bất bình đẳng bằng cách tiếp tục tạo ra các kết quả không chính xác, dẫn đến việc gia tăng bất lợi đối với các nhóm dân cư bị thiệt thòi. Các hệ thống AI có thể bị sử dụng cho bạo lực giới, khiến các cá nhân phải rút lui khỏi không gian công cộng. Các quyết định mang tính phân biệt đối xử hoặc không công bằng có thể không được giải quyết và lặp lại trong tương lai. Nạn nhân có thể không thể tiếp cận biện pháp khắc phục sau khi quyền con người của họ bị vi phạm hoặc lạm dụng.

1.3. Thách thức đối với quyền con người liên quan đến AI

Khó xác định trách nhiệm: Xác định ai chịu trách nhiệm cho các quyết định do các hệ thống AI đưa ra là một thách thức, đặc biệt là những hệ thống liên quan đến quyết định tự động; không rõ trách nhiệm thuộc về các nhà phát triển, nhà cung cấp dữ liệu hay cơ quan, tổ chức ứng dụng AI. Các hệ thống AI, đặc biệt là những hệ thống sử dụng các mô hình phức tạp như học sâu (deep learning) có quá trình ra quyết định thiếu rõ ràng. Điều này có thể dẫn đến khó thực thi các biện pháp bảo vệ pháp lý, khắc phục đối với vi phạm quyền con người.

Sai lệch ngữ cảnh: Có xu hướng áp dụng các hệ thống AI phổ quát, không điều chỉnh chúng cho các ngữ cảnh pháp lý, văn hóa hoặc xã hội cụ thể mà chúng được sử dụng; không hiểu sâu về bối cảnh cụ thể hoặc nhu cầu và giá trị cụ thể của các cộng đồng. Ví dụ, một hệ thống AI được thiết kế trong một bối cảnh văn hóa nhất định có thể không tôn trọng hoặc hỗ trợ các quyền được coi trọng trong một bối cảnh khác. Cách tiếp cận này có thể làm suy yếu quyền con người do không đáp ứng nhu cầu đa dạng của các nhóm dân cư khác nhau.

Khoảng trống về quy định và đạo đức: Sự phát triển của AI đã đi nhanh hơn các quy định và hướng dẫn đạo đức về AI. Khoảng cách này có thể làm cho các hệ thống AI được triển khai mà không có sự giám sát đầy đủ, làm tăng nguy cơ vi phạm quyền con người. Việt Nam vẫn đang trong những bước đầu xây dựng khung pháp lý và quy tắc ứng xử đối với AI; việc tích hợp các chuẩn mực quốc tế về quyền con người để giảm thiểu các rủi ro liên quan đến AI vẫn chưa được đề cập nhiều. Hơn nữa, cơ chế khiếu nại, khiếu kiện, giải quyết tranh chấp nói chung, cũng như khiếu kiện về quyền con người là một khoảng trống lớn ở Việt Nam, trong đó có liên quan đến giải quyết các vi phạm từ việc triển khai, ứng dụng AI gây ra.

II. Kinh nghiệm quốc tế về chính sách Kiểm soát và hạn chế rủi ro cuả AI

2.1. Hoa Kỳ và tiếp cận thúc đẩy đổi mới sáng tạo

Đổi mới sáng tạo được nhìn nhận như một thế mạnh và yếu tố then chốt tạo nên sức mạnh của Hoa Kỳ. Từ khi bắt đầu làn sóng AI Internet[4] vào thập niên 90, các nhà lập pháp Hoa Kỳ đã xây dựng các quy định pháp lý quan trọng thể hiện trong Đạo luật Viễn thông (Telecommunications Act) năm 1996 và Đạo luật bản quyền thiên niên kỷ kỹ thuật số (Digital Millennium Copyright Act) năm 1998, góp phần làm nền tảng cho sự phát triển mạnh mẽ của các công nghệ như Google, Microsoft, Meta, Youtube sau này. Cụ thể, các doanh nghiệp cung cấp dịch vụ số trung gian sẽ không bị đối xử như các nhà xuất bản hay người cung cấp tin tức; và trong lĩnh vực bản quyền, các doanh nghiệp này được loại trừ trách nhiệm liên đới với hành vi xâm phạm bản quyền của người dùng khi đã thực hiện chặn/gỡ theo khiếu nại của chủ thể quyền.

Đến hiện tại, nhánh lập pháp Hoa Kỳ chưa có thêm bước tiến mới nào trong việc xác định khuôn khổ pháp lý đối với AI. Còn nhánh hành pháp dường như tiếp tục theo đuổi đổi mới sáng tạo thể hiện ở cách tiếp cận điều tiết "mềm" - thúc đẩy các cam kết tự nguyện trong ngành thông qua các hướng dẫn. Chẳng hạn, Bản hướng dẫn thiết kế, sử dụng, triển khai AI vì các quyền của công dân Mỹ (Blueprint for an AI Bill of Rights) được Nhà Trắng công bố vào tháng 10/2022. Sang tháng 8/2023, Sắc lệnh về phát triển và sử dụng AI an toàn, bảo mật, đáng tin cậy (Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence) được Tổng thống Biden ban hành. Sau đó, Viện Tiêu chuẩn và Công nghệ quốc gia thuộc Bộ Thương mại Hoa Kỳ mới đây đã bắt đầu lấy ý kiến công chúng để làm cơ sở xây dựng hướng dẫn đánh giá, phân loại AI và phát triển các tiêu chuẩn dựa trên đồng thuận chung.

2.2. EU và tiếp cận Ưu tiên quản lý rủi ro

Với cách tiếp cận nhất quán quản lý công nghệ số dựa trên rủi ro nhằm bảo vệ tự do và phẩm giá con người, EU đã liên tục ban hành các quy định mới để điều chỉnh các vấn đề phát sinh đi liền với sự phát triển của công nghệ. Có thể kể đến như Quy định chung về Bảo vệ dữ liệu (GDPR) nhằm giải quyết vấn đề về quyền riêng tư trên không gian mạng, Đạo luật về Dịch vụ số (Digital Services Act) nhằm giải quyết vấn đề về an toàn không gian mạng và quyền lợi người dùng hay Đạo luật về Thị trường số (Digital Market Act) nhằm giải quyết vấn đề về cạnh tranh lành mạnh trong cung cấp các dịch vụ số. Mới đây nhất, EU đã đạt được thỏa thuận chính trị trong nội khối, giữa Nghị viện EU và Hội đồng EU, tiến tới chính thức thông qua đạo luật về AI. Đạo luật này cùng với các văn bản như GDPR, Đạo luật Dịch vụ Kỹ thuật số (DSA), Đạo luật Thị trường kỹ thuật số (DMA) sẽ trở thành bộ công cụ chính sách giúp EU thích ứng với bối cảnh công nghệ và đưa EU trở thành khu vực tiên phong trong thiết lập các tiêu chuẩn pháp lý mới đối với công nghệ số trên toàn cầu.

[4] Tiến trình bốn làn sóng AI theo nhà khoa học máy tính Kai-Fu Lee: AI Internet - AI kinh doanh - AI Nhận thức - AI Tự quản

Tại EU, AI được tiếp cận dựa trên thang rủi ro đối với con người với 04 mức độ từ cao xuống thấp gồm: rủi ro không thể chấp nhận (unacceptable risk), rủi ro cao (high risk), rủi ro giới hạn (limited risk) và rủi ro tối thiểu (minimal risk). Dựa trên mức độ rủi ro này, các nhà lập pháp thiết kế những quy định tương ứng để đảm bảo AI trở nên đáng tin cậy, an toàn và trong tầm kiểm soát đối với con người. Cụ thể, AI gây ra rủi ro không thể chấp nhận là các hệ thống AI đe dọa rõ ràng đến các quyền cơ bản, phá hoại tự do ý chí của cá nhân chẳng hạn như hệ thống AI cho phép chính phủ hoặc doanh nghiệp chấm điểm xã hội (social scoring), sẽ bị cấm ở EU. AI gây ra rủi ro cao như hệ thống AI được dùng trong tuyển dụng nhân sự, thực thi pháp luật sẽ phải đáp ứng các yêu cầu nghiêm ngặt về chất lượng dữ liệu, ghi nhật kí hệ thống, vận hành với sự giám sát của con người. AI gây ra rủi ro giới hạn như chatbot sẽ phải được gắn nhãn để người dùng nhận biết được họ đang tương tác với máy móc. Và AI có rủi ro tối thiểu như hệ thống gợi ý hoặc bộ lọc thư rác sẽ không phải đáp ứng các yêu cầu cứng trong luật. Thay vào đó, trên cơ sở tự nguyện, các doanh nghiệp vận hành hệ thống AI rủi ro tối thiểu có thể áp dụng các bộ quy tắc ứng xử (code of conduct) cho các hệ thống AI này.

2.3. Trung Quốc: Coi trọng đồng thời an ninh và phát triển AI

Luật An ninh mạng, Luật Bảo mật dữ liệu, Luật Bảo vệ thông tin cá nhân do Cơ quan quản lý không gian mạng Trung Quốc (Cyberspace Administration of China - CAC) xây dựng và thực thi là ba trong số nhiều văn bản cho thấy nước này xây dựng các biện pháp quản lý công nghệ số dựa trên bảo đảm an ninh. Chẳng hạn, yêu cầu dữ liệu phải được lưu trữ tại Trung Quốc đại lục (data localization), thực hiện cấp phép đối với hoạt động chuyển dữ liệu cá nhân ra khỏi lãnh thổ Trung Quốc, để đảm bảo "tài nguyên số" nằm trong chủ quyền quốc gia.

Tiếp nối cách tiếp cận này, tháng 4/2023, CAC đã xây dựng Các biện pháp tạm thời quản lý dịch vụ AI tạo sinh (Interim Measures for the Management of Generative Artificial Intelligence Services) có hiệu lực từ ngày 15/8/2023. Văn bản này khẳng định nguyên tắc quản lý AI dịch vụ AI của Trung Quốc là coi trọng an ninh và phát triển như nhau (**国家**坚持发展和安全并重). Các hoạt động triển khai dịch vụ AI tạo sinh được Đảng và Nhà nước theo dõi thận trọng để đảm không làm xói mòn các giá trị cốt lõi của chủ nghĩa xã hội đặc sắc Trung Quốc. Sự thận trọng đối với dịch vụ AI tạo sinh tương tự với cách tiếp cận với thương mại điện tử, công nghệ tài chính, tiền mã hóa trước đây. Cách tiếp cận này dường như xuất phát từ việc chưa nhận định rõ ràng rủi ro và cơ hội mà công nghệ mới tạo ra, do đó, một mặt chính quyền vẫn để thị trường phát triển, một mặt muốn đảm bảo thị trường phát triển đúng định hướng chủ nghĩa xã hội đặc sắc Trung Quốc.

III. Khuyến nghị chính sách cho Việt Nam

3.1. Quản lý rủi ro cần chú trọng tính cân bằng giữa kiểm soát và phát triển

Cân bằng giữa kiểm soát rủi ro với phát triển AI: Cần xem xét tác động cụ thể của AI đối với lợi ích của người dùng cá nhân và xã hội trong toàn bộ quá trình thiết kế, phát triển, triển khai, ứng dụng các hệ thống AI như đã trình bày trong bảng ở trên[5]. Điều này nhằm đảm bảo rằng AI không chỉ hoạt động hiệu quả về mặt kỹ thuật, mà còn phù hợp với các giá trị xã hội và thúc đẩy, bảo vệ các quyền con người.

Mặt khác, kiểm soát, giảm thiểu rủi ro xã hội và bảo vệ quyền con người nhưng không làm cản trở hay đình trệ các tiến bộ và cải cách sáng tạo. Các biện pháp, quy định phải cân bằng được yêu cầu vừa tạo ra môi trường và độ linh hoạt cho sự phát triển AI, đồng thời kiểm soát, giảm thiểu các mối nguy cơ, rủi ro với xã hội. Có thể đạt mục tiêu này bằng cách đánh giá, điều chỉnh AI theo các mức độ rủi ro để có phương án phù hợp với mức độ rủi ro đã được xác định.

3.2. Tiếp cận toàn diện: Các bên liên quan đều tham gia và có vai trò trong thúc đẩy an toàn AI

Các bên liên quan trong xã hội đa dạng cần được tham gia từ đầu trong vòng đời của AI; không chỉ các chuyên gia kỹ thuật, doanh nghiệp, mà còn cả các chuyên gia pháp lý, các chuyên gia về quyền con người, đại diện từ cộng đồng, các tổ chức xã hội, và các nhóm khác có thể bị ảnh hưởng bởi hệ thống AI. Trong đó, cần đặc biệt tham vấn ý kiến của các cơ sở nghiên cứu, thực hành về quyền con người, cụ thể ở Việt Nam là Viện Nghiên cứu quyền con người thuộc Học viện Chính trị quốc gia Hồ Chí Minh, Trung tâm Nghiên cứu quyền con người thuộc Trường Luật, Đại học quốc gia Hà Nội v.v…Bên cạnh đó, nâng cao năng lực về công nghệ số, trong đó có AI, nhận thức, kỹ năng sử dụng AI an toàn cho tất cả các chủ thể có thể góp phần đảm bảo quyền con người.

3.3. Thiết kế Quy định về AI trong dự thảo luật Công nghiệp Công nghệ số

Nên chọn cách xây dựng quy định pháp lý theo hướng ‘nới lỏng’, khuyến khích tổ chức, doanh nghiệp, cá nhân thực hiện cơ chế tuân thủ tự nguyện qua xây dựng và thực hiện các bộ quy tắc ứng xử; hoặc tuân thủ các bộ tiêu chuẩn công nghệ do ngành công nghệ đặt ra. Cụ thể, nội dung về AI trong dự thảo Luật Công nghiệp công nghệ số nên đưa ra các nguyên tắc về phát triển, ứng dụng AI an toàn; quy định trách nhiệm về việc xây dựng, thực hiện các bộ quy tắc ứng xử về đạo đức và an toàn AI. Luật chưa nên đưa ra các nghĩa vụ pháp lý mang tính bắt buộc phải thực thi cho các chủ thể.

---

[5] Evgeni Aizenberg and Jeroen van den Hoven, Designing for Human Rights in AI; Bureau of Cyberspace and Digital Policy, Risk Management Profile for Artificial Intelligence and Human Rights, 7/2024, https://www.state.gov/risk-management-profile-for-ai-and-human-rights/ .