Quản trị Quá trình chuyển đổi sang Trí tuệ Nhân tạo Tổng quát (Artificial General Intelligence AGI) Những cân nhắc cấp bách đối với Đại hội đồng Liên Hợp Quốc

Báo cáo cho Hội đồng các Chủ tịch của Đại hội đồng Liên Hợp Quốc (UNCPGA)

Tóm tắt điều hành

Các hệ thống trí tuệ nhân tạo (AI) đang tiến nhanh tới ngưỡng của trí tuệ nhân tạo tổng quát (AGI) —đặc trưng bởi khả năng thực hiện các nhiệm vụ nhận thức đa dạng với hiệu quả ngang bằng hoặc vượt qua trí thông minh của con người. Với các đầu tư tài chính lớn nhất trong lịch sử đang thúc đẩy những nỗ lực nghiên cứu và phát triển chưa từng có, các nhà lãnh đạo ngành và chuyên gia dự báo rằng AGI có thể xuất hiện ngay trong thập kỷ này,[1] mang lại những lợi ích vượt bậc cho nhân loại. Trong số những lợi ích đó, AGI có thể thúc đẩy nhanh chóng các khám phá khoa học liên quan đến sức khỏe cộng đồng, chuyển đổi nhiều ngành công nghiệp và nâng cao năng suất, đồng thời góp phần hiện thực hóa các Mục tiêu Phát triển Bền vững (SDGs).

Tuy nhiên, AGI cũng có thể tạo ra những rủi ro đặc thù và thậm chí thảm hoạ tiềm tàng. Khác với AI truyền thống, AGI có thể tự động thực thi các hành động gây hại nằm ngoài tầm kiểm soát của con người, dẫn đến những hậu quả không thể đảo ngược được, bao gồm các mối đe doạ từ các hệ thống vũ khí tiên tiến và các lỗ hổng trong hạ tầng thiết yếu. Chúng ta phải bảo đảm rằng những rủi ro này được giảm thiểu nếu muốn khai thác được những lợi ích vượt bậc mà AGI mang lại.

Để ứng phó hiệu quả với những thách thức toàn cầu này, cần có hành động quốc tế khẩn cấp và phối hợp dưới sự hỗ trợ của Liên Hợp Quốc. Những hành động này nên được khởi xướng thông qua một phiên họp đặc biệt của Đại hội đồng Liên Hợp Quốc về AGI nhằm thảo luận về các lợi ích và rủi ro của AGI, cũng như về khả năng thiết lập các cơ chế toàn cầu như: một đài quan sát quốc tế về AGI, hệ thống chứng nhận AGI an toàn và đáng tin cậy, một Công ước Liên Hợp Quốc về AGI, và một cơ quan quốc tế chuyên trách về AGI. Nếu không có cơ chế quản trị toàn cầu chủ động, sự cạnh tranh giữa các quốc gia và tập đoàn sẽ đẩy nhanh việc phát triển AGI đầy rủi ro, làm suy yếu các giao thức an ninh và gia tăng căng thẳng địa chính trị. Hành động quốc tế có phối hợp có thể ngăn ngừa những hệ quả này, thúc đẩy việc phát triển và sử dụng AGI một cách an toàn, bảo đảm phân phối lợi ích một cách công bằng, và duy trì sự ổn định toàn cầu.

I. Giới thiệu

[1] METR: Đo lường khả năng của AI trong việc thực hiện các nhiệm vụ kéo dài (Measuring AI Ability to Compete Long Tasks) https://arxiv.org/abs/2503.14499

Tốc độ phát triển của trí tuệ nhân tạo (AI) đã diễn ra nhanh chóng trong những năm và tháng gần đây,[2] và có thể sẽ còn tăng tốc hơn nữa — một phần do các công ty AI đang đầu tư những khoản tiền khổng lồ để phát triển các tác nhân AI (AI agents) ngày càng có năng lực và tính tự chủ cao hơn, và một phần do việc sử dụng ngày càng nhiều các mô hình AI mạnh nhất để thúc đẩy chính quá trình nghiên cứu về chính AI,[3] Nhiều người kỳ vọng rằng những tiến bộ vượt bậc trong năng lực của trí tuệ nhân tạo sẽ dẫn đến sự ra đời của "Trí tuệ Nhân tạo Tổng quát" (AGI): các hệ thống AI có khả năng ngang bằng hoặc vượt qua con người trong hầu hết các nhiệm vụ nhận thức.

Mặc dù còn có những ý kiến khác nhau về thời điểm AGI sẽ xuất hiện, tất cả các chuyên gia trong Hội đồng này đều tin rằng AGI có thể sẽ được phát triển ngay trong thập kỷ này. Các công ty AI đang cam kết đầu tư hàng trăm tỷ đô la để đạt được AGI trong thời gian sớm, biến đây trở thành nỗ lực nghiên cứu và phát triển (R&D) lớn nhất trong lịch sử loài người. Khu vực tư nhân cần có trách nhiệm phát triển công nghệ theo hướng an toàn hơn và cần có các cơ chế khuyến khích để thực hiện điều đó; tuy nhiên, cuộc chạy đua cạnh tranh nhằm giành vị trí tiên phong trong việc đạt được AGI đang khiến các công ty dồn toàn bộ nguồn lực vào việc nâng cao khả năng hệ thống, thay vì tập trung vào yếu tố an toàn, với mục tiêu "giành chiến thắng cuộc đua".

Những rủi ro hiện tại liên quan đến AI chủ yếu bắt nguồn từ việc con người sử dụng công nghệ sai mục đích. Tuy nhiên, AGI đặt ra một loại rủi ro khác biệt căn bản, vì các mối đe dọa tiềm tàng của nó vượt ra ngoài việc con người sử dụng sai. AGI có thể tự động tạo ra và thực hiện các kế hoạch với hậu quả thảm khốc, vượt qua khả năng của con người trong việc nhận diện, phân tích và ứng phó với các mối đe dọa mới và những biến động chưa từng có tiền lệ.4 Khi kết hợp với khuynh hướng tự bảo toàn5 được quan sát thấy gần đây ở các hệ thống AI tiên tiến, điều này có thể dẫn đến những tình huống mà AGI trở nên không thể kiểm soát được.

Đây cần được xem là một mối quan ngại chung của toàn thể cộng đồng quốc tế. Các rủi ro liên quan đến AGI không chỉ giới hạn trong phạm vi một ngành hay một quốc gia, mà có hệ lụy toàn cầu, bất kể chúng có khởi nguồn từ đâu. Đảm bảo sự tích hợp an toàn và hài hòa của AGI đòi hỏi không chỉ nỗ lực từ quốc gia hoặc khu vực tư nhân mà còn cần đến cơ chế quản trị quốc tế chủ động, được dẫn dắt bởi Liên Hợp Quốc. Liên Hợp Quốc có vị thế đặc biệt để thúc đẩy việc đạt được một thỏa thuận khoa học về các rủi ro và các chiến lược giảm thiểu rủi ro, xây dựng đồng thuận chính trị xung quanh cách tiếp cận chung trong việc giảm thiểu rủi ro, điều phối chính sách, thúc đẩy các tiêu chuẩn hoặc các cơ chế bảo vệ, ứng phó với các tình huống khẩn cấp, cũng như thực hiện hoặc điều phối các nghiên cứu chung về an toàn và an ninh.

---

[2] Xem Báo cáo An toàn AI quốc tế, Bengio và cộng sự 2025.
[3] https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.pdf
[4] "Claude 3.7 (thường) biết khi nào nó đang tham gia các đánh giá về sự tương thích (alignment evaluations)" https://www.apolloresearch.ai/blog/claude-sonnet-37-often-knows-when-its-in-alignment-evaluations
[5] Xem Meinke và cộng sự 2024, Các mô hình tiên phong có khả năng lên kế hoạch có chủ đích trong ngữ cảnh, https://arxiv.org/abs/2412.04984.

Nếu không có cơ chế quản trị toàn cầu, những tiềm năng đột phá của trí tuệ nhân tạo tổng quát (AGI) trong việc giải quyết các thách thức toàn cầu có thể sẽ không được khai thác đầy đủ hoặc có nguy cơ bị định hướng sai lệch. Hơn nữa, sự phối hợp ở quy mô toàn cầu là yếu tố then chốt để ứng phó với các mối đe dọa mang tính thảm họa toàn cầu mà AGI được dự báo có thể gây ra. Thật khó để tưởng tượng rằng sự phối hợp này có thể đạt được trên phạm vi toàn cầu mà thiếu vai trò lãnh đạo tích cực từ Liên Hợp Quốc.

I. Sự cấp thiết của hành động bởi Đại hội đồng Liên Hợp Quốc về quản trị trí tuệ nhân tạo tổng quát và những hệ lụy có thể xảy ra nếu không có hành động kịp thời

Trong bối cảnh địa chính trị phức tạp và hiện vẫn chưa có các chuẩn mực quốc tế mang tính ràng buộc và thống nhất, việc chạy đua phát triển trí tuệ nhân tạo tổng quát (AGI) không đi kèm với các biện pháp an toàn đầy đủ đang làm gia tăng nguy cơ xảy ra sự cố, bị lạm dụng, vũ khí hóa, và thậm chí dẫn đến những thất bại mang tính tồn vong.[6] Các quốc gia và tập đoàn đang đặt ưu tiên vào tốc độ hơn là an toàn, khiến cho các khung quản trị cấp quốc gia bị suy yếu và các giao thức bảo đảm an toàn bị xem là thứ yếu so với lợi ích kinh tế hoặc quân sự. Trong bối cảnh nhiều dạng AGI từ cả chính phủ và doanh nghiệp có thể xuất hiện trước khi thập kỷ này kết thúc, trong khi việc thiết lập các hệ thống quản trị ở cấp quốc gia và quốc tế sẽ cần nhiều năm, việc khẩn trương khởi động các thủ tục cần thiết là điều cấp bách để ngăn chặn những hệ quả sau đây:

1. **Những hệ quả không thể đảo ngược**— Một khi trí tuệ nhân tạo tổng quát (AGI) được phát triển, tác động của nó có thể không thể đảo ngược. Nhiều dạng AI tiên tiến hiện nay đã thể hiện hành vi đánh lừa và tự bảo vệ, cùng với xu hướng phát triển các hệ thống AI ngày càng tự chủ, có khả năng tương tác, tự cải thiện và được tích hợp vào các hạ tầng trọng yếu, quỹ đạo phát triển và tác động của AGI có thể dẫn đến việc không thể kiểm soát được. Nếu điều đó xảy ra, nhân loại có thể không còn khả năng quay trở lại trạng thái kiểm soát tin cậy bởi con người. Việc thiết lập cơ chế quản trị chủ động là hết sức cần thiết để bảo đảm rằng AGI không vượt qua các "lằn ranh đỏ" của chúng ta,[7] dẫn đến những hệ thống vận hành vượt ngoài khả năng con người giành lại quyền điều kiển.

2. **Vũ khí hủy diệt hàng loạt** —AGI có thể tạo điều kiện cho một số quốc gia và chủ thể phi nhà nước có ý đồ xấu phát triển vũ khí hóa học, sinh học, phóng xạ và hạt nhân. Hơn nữa, các đội hình vũ khí sát thương tự động do AGI điều khiển với quy mô lớn có thể tự bản thân cấu thành một loại vũ khí hủy diệt hàng loạt mới.

[6] Phản hồi của OpenAI đối với Kế hoạch Hành động về Trí tuệ Nhân tạo của Văn phòng Chính sách Khoa học và Công nghệ Hoa Kỳ https://cdn.openai.com/global-affairs/ostp-rfi/ec680b75-d539-4653-b297-8bcf6e5f7686/openai-response-ostp-nsf -rfi-notice-request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan.pdf

[7] Đối thoại quốc tế về An toàn AI (2024): https://idais.ai/dialogue/idais-beijing/

3. **Lỗ hổng trong cơ sở hạ tầng thiết yếu**—Các hệ thống hạ tầng quốc gia trọng yếu (như mạng lưới điện, hệ thống tài chính, mạng lưới giao thông, thông tin liên lạc và hệ thống y tế) có thể trở thành mục tiêu của các cuộc tấn công mạng quy mô lớn do AGI trực tiếp phát động hoặc hỗ trợ. Nếu không có cơ chế răn đe hiệu quả ở cấp quốc gia và sự phối hợp chặt chẽ ở cấp quốc tế, các chủ thể phi nhà nước có ý đồ xấu — từ các tổ chức khủng bố đến các mạng lưới tội phạm có tổ chức xuyên quốc gia — có thể tiến hành các cuộc tấn công ở quy mô lớn chưa từng có.

4. **Tập trung quyền lực, bất bình đẳng toàn cầu và bất ổn —**Việc phát triển và sử dụng AGI một cách thiếu kiểm soát có thể làm trầm trọng thêm tình trạng chênh lệch về của cải và quyền lực ở quy mô chưa từng có. Nếu AGI chỉ nằm trong tay một số quốc gia, tập đoàn hoặc nhóm tinh hoa, điều này có thể củng cố thế độc quyền kinh tế và dẫn đến sự thống trị toàn cầu về tri thức, đổi mới sáng tạo và năng lực sản xuất công nghiệp. Hệ quả có thể bao gồm tình trạng thất nghiệp trên diện rộng, suy giảm quyền năng của người dân ảnh hưởng đến nền tảng pháp lý, mất quyền riêng tư và sụp đổ lòng tin vào các thể chế, tri thức khoa học cũng như hệ thống quản trị. AGI có thể làm suy yếu các thể chế dân chủ thông qua các hình thức thuyết phục, thao túng và tuyên truyền do AI tạo ra, đồng thời làm gia tăng bất ổn địa chính trị theo cách mà làm tăng thêm các lỗ hổng hệ thống. Việc thiếu phối hợp giữa các quốc gia có thể dẫn đến xung đột về tài nguyên, năng lực hoặc quyền kiểm soát AGI, với nguy cơ leo thang thành đối đầu quân sự. AGI cũng sẽ đặt ra thách thức lớn đối với các khuôn khổ pháp lý hiện hành: nhiều vấn đề mới và phức tạp về sở hữu trí tuệ, trách nhiệm pháp lý, nhân quyền và chủ quyền có thể vượt quá khả năng ứng phó của cả hệ thống pháp luật trong nước lẫn quốc tế..

5. **Rủi ro tồn vong—** AGI có thể bị lạm dụng để gây tổn hại trên diện rộng hoặc được phát triển theo cách không phù hợp với các giá trị cốt lõi của con người. AGI thậm chí có thể vận hành một cách tự chủ vượt ngoài tầm giám sát của con người, tự thiết lập các mục tiêu riêng dựa trên cơ chế tự bảo vệ – điều đã được ghi nhận ở một số hệ thống AI tiên tiến hiện nay. AGI cũng có thể chủ động tìm kiếm quyền lực như một phương tiện nhằm đảm bảo nó có thể thực hiện bất kỳ mục tiêu nào mà nó đặt ra, bất chấp sự can thiệp của con người. Các chính phủ, chuyên gia hàng đầu và các công ty đang phát triển AGI đều đã lên tiếng cảnh báo rằng những xu hướng hiện tại có thể dẫn đến các kịch bản trong đó hệ thống AGI tìm cách chế ngự con người. Đây không còn là những giả thuyết khoa học viễn tưởng xa vời trong tương lai — nhiều chuyên gia hàng đầu nhận định rằng các rủi ro này hoàn toàn có thể xảy ra ngay trong thập kỷ này, và những dấu hiệu ban đầu đã bắt đầu xuất hiện. Hơn nữa, cho đến nay, các nhà phát triển AI hàng đầu vẫn chưa đưa ra được một giải pháp khả thi nào có thể ngăn ngừa những rủi ro này với độ tin cậy cao.

6. **Mất đi những lợi ích vượt bậc trong tương lai cho toàn nhân loại —** Nếu được quản trị đúng cách, trí tuệ nhân tạo tổng quát (AGI) hứa hẹn mang lại những tiến bộ vượt bậc trong mọi lĩnh vực, cho tất cả mọi người — từ y học cá nhân hóa, chữa trị ung thư, tái tạo tế bào, đến hệ thống học tập cá nhân hóa, chấm dứt đói nghèo, ứng phó biến đổi khí hậu, và thúc đẩy các khám phá khoa học với lợi ích ngoài sức tưởng tượng. Để thực hiện hoá một tương lai tươi sáng như vậy cho toàn nhân loại, cần có một cơ chế quản trị toàn cầu, bắt đầu bằng việc nâng cao nhận thức chung trên toàn cầu về cả rủi ro và cơ hội mà AGI mang lại. Liên Hợp Quốc đóng vai trò trung tâm trong sứ mệnh này.

Ⅱ. Mục tiêu của Liên Hợp Quốc trong việc quản trị quá trình chuyển đổi sang trí tuệ nhân tạo tổng quát (AGI)

Xét đến khả năng AGI có thể được phát triển ngay trong thập kỷ này, việc xây dựng các cấu trúc quản trị vững chắc là điều cấp thiết cả về mặt khoa học lẫn đạo đức để chuẩn bị sẵn sàng cho cả những lợi ích vượt trội và những rủi ro to lớn mà công nghệ này có thể tạo ra.

Mục đích của Liên Hợp Quốc trong việc quản trị quá trình chuyển đổi sang trí tuệ nhân tạo tổng quát (AGI) là nhằm bảo đảm rằng việc phát triển và sử dụng AGI phù hợp với các giá trị phổ quát của con người, an ninh quốc tế và phát triển toàn cầu. Điều này bao gồm: 1) Thúc đẩy nghiên cứu về khả năng điều chỉnh và kiểm soát AI để xác định các phương pháp kỹ thuật cho việc định hướng và/hoặc kiểm soát các hệ thống AI ngày càng có năng lực vượt trội; 2) Đưa ra định hướng cho quá trình phát triển AGI — xây dựng các khuôn khổ nhằm bảo đảm rằng AGI được phát triển một cách có trách nhiệm, với các biện pháp bảo mật vững chắc, tính minh bạch cao và phù hợp với các giá trị cốt lõi của con người; 3) Xây dựng các cơ chế quản trị cho việc triển khai và ứng dụng AGI — ngăn chặn lạm dụng, bảo đảm tiếp cận công bằng giữa các quốc gia và cộng đồng, tối đa hóa lợi ích cho nhân loại đồng thời giảm thiểu các rủi ro; 4) Thúc đẩy tầm nhìn tương lai về một AGI vì lợi ích chung — xây dựng các khuôn khổ mới phục vụ phát triển xã hội, môi trường và kinh tế; 5) Thiết lập một nền tảng trung lập và không loại trừ ai cho hợp tác quốc tế — đặt ra các tiêu chuẩn toàn cầu, xây dựng khuôn khổ pháp lý quốc tế, và tạo lập các cơ chế khuyến khích tuân thủ; qua đó, củng cố lòng tin giữa các quốc gia nhằm bảo đảm tiếp cận công bằng trên toàn cầu với các lợi ích mà AGI mang lại.

Ⅲ. Phiên họp Đại hội đồng Liên Hợp quốc về những vấn đề then chốt cần xem xét

Một trong những thách thức lớn nhất trong quản trị AGI là sự bất định cao liên quan đến quá trình phát triển công nghệ trong tương lai. Điều này gây khó khăn trong việc dự báo chính xác các lợi ích và rủi ro tiềm ẩn. Do đó, cần thiết phải xây dựng một khuôn khổ ứng phó toàn diện và bao quát nhằm chủ động lường trước và giảm thiểu các mối đe dọa có thể xảy ra, đồng thời tăng cường khả năng hiện thực hóa các lợi ích

tiềm năng. Liên Hợp Quốc có thể đóng vai trò điều phối quốc tế then chốt đối với quá trình phát triển và sử dụng AGI. Đặc biệt quan trọng là việc bảo đảm sự tham gia đầy đủ của tất cả các quốc gia trong tiến trình này và thu hẹp chia rẽ địa chính trị; hiện nay, chỉ Liên Hợp Quốc là tổ chức có đủ vị thế phù hợp để đảm nhận vai trò này một cách hiệu quả. Các nội dung sau đây cần được xem xét trong khuôn khổ phiên họp chuyên đề của Đại hội đồng Liên Hợp Quốc về AGI:

A. Đài quan sát toàn cầu về AGI

Cần thành lập một Đài quan sát toàn cầu về AGI để theo dõi tiến độ nghiên cứu và phát triển liên quan đến AGI, và đưa ra cảnh báo sớm về các vấn đề an ninh AI cho các Quốc gia Thành viên. Đài quan sát này nên tận dụng chuyên môn và nguồn lực từ các sáng kiến hiện có của Liên Hợp Quốc, chẳng hạn như Ban Khoa học Quốc tế Độc lập về AI được thành lập trong khuôn khổ Thoả thuận Kỹ thuật Số Toàn cầu (Global Digital Compact) và Bộ công cụ đánh giá mức độ sẵn sàng do UNESCO xây dựng.

B. Hệ thống quốc tế về các thông lệ tốt và chứng nhận cho các AGI an toàn và đáng tin cậy

Xét đến khả năng AGI có thể được phát triển ngay trong thập kỷ này, việc xây dựng các cấu trúc quản trị vững chắc là một yêu cầu cấp thiết cả về mặt khoa học lẫn đạo lý, nhằm chuẩn bị sẵn sàng cho cả những lợi ích vượt trội và những rủi ro nghiêm trọng mà công nghệ này có thể mang lại.

C. Công ước Khung của Liên Hợp Quốc về AGI

Cần có một Công ước Khung về AGI[8] để thiết lập các mục tiêu chung và các giao thức linh hoạt nhằm quản lý rủi ro AGI và đảm bảo phân phối lợi ích toàn cầu công bằng. Công ước này cần phân loại rõ ràng các cấp độ rủi ro tương ứng với các hành động quốc tế phù hợp — từ thiết lập tiêu chuẩn, cơ chế cấp phép, đến việc thành lập các cơ sở nghiên cứu chung đối với những hệ thống AGI có mức độ rủi ro cao, cũng như xác định các ranh giới đỏ và cơ chế cảnh báo sớm[9] trong quá trình phát triển AGI. Công ước sẽ tạo nền tảng thể chế có tính thích ứng để bảo đảm tính chính danh toàn cầu, quản trị AGI toàn diện và hiệu quả, từ đó giảm thiểu rủi ro toàn cầu và tối đa hóa thịnh vượng chung từ AGI.

---

[8] Cass-Beggs, Duncan, Stephen Clare, Dawn Dimowo và Zaheed Kara. 2024. "Công ước Khung về Thách thức AI toàn cầu." Trung tâm Đổi mới Quản trị Quốc tế. https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/

[9] Russell, Stuart, Edson Prestes, Mohan Kankanhalli, Jibu Elias, Constanza Gómez Mont, Vilas Dhar, Adrian Weller, Pascale Fung và Karim Beguir, "Ranh giới đỏ AI: Cơ hội và thách thức trong việc thiết lập các giới hạn." Diễn đàn Kinh tế Thế giới, ngày 11 tháng 3 năm 2025. https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/
Karnofsky, Holden. 2024. "Phác thảo các năng lực cảnh báo sớm tiềm năng đối với AI." Quỹ Carnegie vì Hòa bình Quốc tế. Ngày 10 tháng 12 năm 2024. https://carnegieendowment.org/research/2024/12/a-sketch-of- potential-tripwire-capabilities-for-ai?lang=en

D. Nghiên cứu khả thi về việc thành lập một cơ quan AGI thuộc Liên Hợp Quốc

Xét đến các biện pháp cần thiết ở phạm vi lớn để chuẩn bị cho AGI và tính cấp bách của vấn đề, cần có các bước triển khai để nghiên cứu khả thi việc thành lập một cơ quan chuyên trách của Liên Hợp Quốc về AGI —một quy trình được đẩy nhanh là lý tưởng nhất. Một mô hình tương tự như Cơ quan Năng lượng Nguyên tử Quốc tế (IAEA) đã được đề xuất, với nhận thức rằng quản trị AGI phức tạp hơn nhiều so với năng lượng hạt nhân, do đó cần có những cân nhắc đặc thù trong quá trình nghiên cứu khả thi.

Ⅳ. Những khuyến nghị này đóng góp vào việc thực hiện Hiệp ước Liên Hợp Quốc Cho Tương Lai cũng như các sáng kiến khác của Liên Hợp Quốc

Nhiều sáng kiến của Liên Hợp Quốc kêu gọi phát triển AI an toàn, bảo mật và đáng tin cậy. Trong số đó, các nghị quyết của Đại hội đồng Liên Hợp Quốc về AI–A/78/L.49, A/78/L.86 và A/C.1/79/L.43–cùng với Hiệp ước Liên Hợp Quốc cho Tương lai, Thoả thuận Kỹ thuật Số Toàn cầu và Khuyến nghị của UNESCO về Đạo đức trong Trí tuệ Nhân tạo kêu gọi tăng cường hợp tác quốc tế nhằm phát triển AI vì lợi ích chung cho toàn nhân loại, đồng thời chủ động quản lý các rủi ro toàn cầu.

Các sáng kiến này đã thu hút sự quan tâm của thế giới đến các hình thức trí tuệ nhân tạo hiện nay. Báo cáo này tiếp nối các sáng kiến của Liên Hợp Quốc bằng cách tập trung cụ thể vào sự phát triển của trí tuệ nhân tạo tổng quát trong tương lai gần.

Những cam kết được nêu trong Hiệp ước cho Tương Lai đã được thúc đẩy theo nhiều cách thông qua báo cáo này. Việc tổ chức một phiên họp của Đại Hội đồng Liên Hợp Quốc tập trung vào AGI thể hiện sự hưởng ứng đối với cam kết của Hiệp ước trong việc thúc đẩy đối thoại toàn cầu về quản trị AI. Các khuyến nghị trong báo cáo này về việc xây dựng một Công ước Khung của Liên Hợp Quốc về AGI cũng như tiến hành nghiên cứu khả thi cho việc thành lập một cơ quan AGI trực thuộc Liên Hợp Quốc là những bước đi cụ thể nhằm hiện thực hóa cam kết đó. Đài quan sát mà chúng tôi đề xuất sẽ hỗ trợ cho hoạt động của Ban Cố vấn Khoa học Quốc tế Độc lập về AI sắp được thành lập — một trong những kết quả then chốt của Thoả thuận Kỹ thuật Số Toàn cầu. Cuối cùng, Hệ thống Thực hành Tốt và Chứng nhận Quốc tế cho AGI An toàn và Đáng tin cậy sẽ góp phần tăng cường lòng tin và tính minh bạch theo lời kêu gọi của các Nghị quyết Đại Hội đồng Liên Hợp Quốc, UNESCO và Hiệp ước Vì Tương Lai.

Ⅴ. Kết luận

Việc nâng cao nhận thức của các nhà lãnh đạo quốc gia và quốc tế về lợi ích và rủi ro của AGI trong tương lai— khác biệt về bản chất so với AI hiện tại— là điều hết sức cấp thiết. Đại Hội đồng Liên Hợp Quốc là diễn đàn phù hợp để khởi xướng cuộc thảo luận

toàn cầu này.

Sự phối hợp quốc tế về phát triển và sử dụng AGI là điều cần thiết nhằm bảo đảm rằng nhân loại có thể gặt hái được những lợi ích vượt trội từ AGI, đồng thời bảo vệ quyền con người và an ninh toàn cầu. Trên tinh thần đó, Nhóm Chuyên gia về AGI kiến nghị mạnh mẽ rằng Đại Hội đồng Liên Hợp Quốc khẩn trương hành động để giải quyết các vấn đề nêu trên trong khuôn khổ một phiên họp đặc biệt để xây dựng một khuôn khổ quản trị toàn cầu cho AGI. Nếu không hành động như vậy, rủi ro của việc phát triển và sử dụng AGI một cách thiếu kiểm soát — từ gia tăng bất bình đẳng toàn cầu một cách đột biến cho đến các mối đe dọa mang tính tồn vong — là vô cùng nghiêm trọng. Cách tiếp cận do Liên Hợp Quốc dẫn dắt — bao gồm một đài quan sát toàn cầu, hệ thống chứng nhận quốc tế, một Công ước về AGI của Liên Hợp Quốc, và một cơ quan chuyên trách về AGI — sẽ tăng khả năng AGI được phát triển và ứng dụng vì lợi ích của toàn nhân loại đồng thời giảm thiểu tối đa các rủi ro. Khuôn khổ này cần bảo đảm tính bao trùm không loại trừ ai, minh bạch, và có khả năng thực thi nhằm thúc đẩy lòng tin và hợp tác giữa các quốc gia.

Phụ lục

Điều khoản Tham chiếu: Nhóm Chuyên gia Cấp cao về Trí tuệ Nhân tạo Tổng quát (AGI) cho Hội đồng các Chủ tịch Đại Hội đồng Liên Hợp Quốc (UNCPGA)

Bối cảnh

Tuyên bố Seoul 2024 của Hội đồng Chủ tịch Đại Hội đồng Liên Hợp Quốc (UNCPGA) kêu gọi thành lập một Nhóm Chuyên gia về Trí tuệ Nhân tạo Tổng quát (AGI) để xây dựng khuôn khổ và hướng dẫn chính sách cho Đại Hội đồng Liên Hợp Quốc xem xét trong việc ứng phó những vấn đề cấp bách liên quan đến quá trình chuyển đổi sang AGI.

Công tác này cần được xây dựng trên cơ sở kế thừa, đồng thời tránh trùng lặp với các nỗ lực liên quan đến các giá trị và nguyên tắc về trí tuệ nhân tạo của UNESCO, OECD, G20, G7, Đối tác Toàn cầu về AI (GPAI), Tuyên bố Bletchley, và các khuyến nghị của Nhóm Tư vấn Cấp cao của Tổng Thư ký Liên Hợp Quốc về AI, Thỏa thuận số Toàn cầu của LHQ, Mạng lưới Quốc tế của Viện An toàn AI, Công ước Khung của Hội đồng châu Âu về AI, và hai Nghị quyết của Đại Hội đồng Liên Hợp Quốc về AI. Tuy nhiên, các sáng kiến này chủ yếu tập trung vào những dạng AI hẹp hơn. Hiện vẫn còn thiếu sự quan tâm tương xứng đối với AGI.

AI đã quá quen thuộc với thế giới ngày nay và được sử dụng thường nhưng AGI thì không và hiện vẫn chưa tồn tại. Tuy nhiên, nhiều chuyên gia về AGI cho rằng công nghệ này có thể được phát triển trong vòng từ 1 đến 5 năm tới, và theo thời gian, có thể tiến hóa thành siêu trí tuệ nhân tạo (artificial super intelligence) vượt ngoài khả năng kiểm soát của con người. Hiện chưa có một định nghĩa được thống nhất toàn cầu về trí tuệ nhân tạo tổng quát (AGI), song phần lớn các chuyên gia trong lĩnh vực này đồng thuận rằng AGI sẽ là một hệ thống trí tuệ nhân tạo có tính mục đích tổng quát, có khả năng tự học, tự chỉnh sửa mã lập trình của chính nó, và hành động một cách tự chủ để giải quyết nhiều vấn đề mới bằng những giải pháp sáng tạo, tương đương hoặc vượt trội so với năng lực của con người. Các hệ thống AI hiện nay chưa đạt đến các năng lực này, nhưng xu hướng tiến bộ kỹ thuật hiện tại đang rõ ràng chỉ ra theo hướng đó.

Thỏa thuận Số Toàn cầu của Liên Hợp Quốc (UN Global Digital Compact) kêu gọi tiến hành Đối thoại Toàn cầu về quản trị trí tuệ nhân tạo trong khuôn khổ Liên Hợp Quốc. Các chuyên gia AGI trong khu vực tư nhân đã nhấn mạnh nhu cầu cấp thiết cho một cuộc đối thoại toàn cầu để hiểu rõ hơn về những cơ hội và rủi ro liên quan đến AGI. Một Phiên họp Đặc biệt của Đại Hội đồng Liên Hợp Quốc về AGI có thể là cách nhanh chóng nhất, tiết kiệm chi phí nhất, và có tác động sớm nhất để thúc đẩy cuộc đối thoại mang tính toàn cầu

này.

Mục đích

Triển khai theo Tuyên bố Seoul 2024 của Hội đồng Chủ tịch Đại Hội đồng Liên Hợp Quốc (UNCPGA), đề nghị xây dựng một báo cáo sơ khởi trình Chủ tịch UNCPGA và các Thành viên cho cuộc họp UNCPGA diễn ra từ ngày 8 đến ngày 10 tháng 4 năm 2025 tại Bratislava.

Báo cáo cần xác định rõ những rủi ro, thách thức và cơ hội liên quan đến AGI. Báo cáo cần tập trung nâng cao nhận thức về việc thúc đẩy huy động Đại Hội đồng Liên Hợp Quốc giải quyết vấn đề quản trị AGI một cách có hệ thống hơn. Báo cáo nên tập trung vào AGI – dù hiện chưa đạt được, thay vì các hệ thống AI hẹp đang hiện hành. Đồng thời, báo cáo cần nhấn mạnh tính cấp thiết của việc giải quyết càng sớm càng tốt các vấn đề liên quan đến AGI trong bối cảnh công nghệ này đang phát triển nhanh chóng - có thể đặt ra những rủi ro nghiêm trọng đối với nhân loại cũng như mang lại những lợi ích vượt bậc cho toàn thể loài người.

Báo cáo cũng nên bao gồm cả các thỏa thuận đa phương và các hành động từ khu vực tư nhân để giải quyết những thách thức chưa từng có tiền lệ này. Báo cáo nên phản hồi lời kêu gọi từ các lãnh đạo khu vực tư nhân trong lĩnh vực AGI về việc phối hợp quốc tế và hành động đa phương đối với những gì có thể là thách thức quản trị khó khăn nhất mà nhân loại phải từng đối mặt.

Quy trình

- Triệu tập một nhóm chuyên gia cấp cao (5–8 thành viên) gồm các chuyên gia quốc tế AGI về các mối đe dọa tiềm tàng của AGI đối với nhân loại, cũng như các cơ hội mà AGI có thể mang lại cho lợi ích chung của con người và các vấn đề chính sách liên quan.
- Nhóm chuyên gia về AGI sẽ họp định kỳ trực tuyến bắt đầu từ tháng 1 năm 2025, và hoàn thiện báo cáo sơ khởi để trình bày tại Phiên họp UNCPGA tại Bratislava vào mùa Xuân năm 2025.
- Trên cơ sở các ý kiến đóng góp đối với báo cáo sơ khởi tại Phiên họp UNCPGA ở Bratislava, Nhóm chuyên gia sẽ hoàn thiện báo cáo và trình lên Tổng Thư ký UNCPGA. Nếu được Chủ tịch UNCPGA chấp thuận, báo cáo sẽ được chuyển tiếp đến Chủ tịch Đại hội đồng Liên Hợp Quốc, dự kiến vào ngày 1 tháng 5 năm 2025.

Các thành viên Nhóm Chuyên gia Độc lập Cấp cao về AGI cho Hội đồng Chủ tịch Đại hội đồng Liên Hợp Quốc

Jerome Glenn (Hoa Kỳ) Chủ tịch

Thành viên có quyền biểu quyết thuộc Quản trị tổ chức của IEEE về Trí tuệ nhân tạo; tác giả của tài liệu chương trình Horizon 2025–27 của Liên minh Châu Âu về *AGI: Các vấn đề và Cơ hội;* Giám đốc điều hành của Dự án Thiên niên kỷ (The Millennium Project) và là tác giả của các báo cáo: *"Các vấn đề quản trị quốc tế của quá trình chuyển đổi từ Trí tuệ nhân tạo hẹp sang AGI", "Yêu cầu quản trị toàn cầu đối với AGI", và "Công việc/Công nghệ 2050: Các kịch bản và hành động". Tác giả của cuốn sách Tư duy tương lai: Trí tuệ nhân tạo (1989).*

Renan Araujo (Brazil)

Quản lý nghiên cứu tại Viện Chính sách và Chiến lược về Trí tuệ nhân tạo (Institute for AI Policy and Strategy – IAPS), tập trung vào quản lý rủi ro liên quan đến phát triển AGI. Hiện ông đang dẫn dắt các hoạt động của IAPS về quản trị AGI ở cấp độ quốc tế. Ông là Nghiên cứu viên của Phòng Thí nghiệm Chính sách Trung Quốc tại Đại học Oxford, là luật sư, đồng sáng lập Sáng kiến Condor (kết nối sinh viên Brazil với các cơ hội đẳng cấp thế giới nhằm định hình nghiên cứu và chính sách AI), và từng làm việc trong các chương trình quản trị AI tại tổ chức Rethink Priorities và Viện Luật và Trí tuệ nhân tạo (Institute for Law and AI).

Yoshua Bengio (Canada)

Giáo sư khoa học máy tính tại Đại học Montréal; Chủ tịch Nhóm Cố vấn về An toàn và Bảo mật AI cho chính phủ Canada; Chủ tịch Báo cáo Quốc tế về An toàn AI do 30 quốc gia cùng với Liên Hợp Quốc, OECD và Liên minh Châu Âu ủy nhiệm. Giám đốc khoa học của Mila – Viện Trí tuệ nhân tạo Québec; Thành viên Hội đồng Cố vấn Khoa học của Tổng Thư ký Liên Hợp Quốc về Những Đột Phá trong Khoa học và Công nghệ; là người nhận Giải thưởng Turing và hiện là nhà khoa học máy tính được trích dẫn nhiều nhất trên toàn cầu.

Joon Ho Kwak (Hàn Quốc)

Cố vấn kỹ thuật của Viện An toàn AI Hàn Quốc; đóng vai trò chủ chốt trong việc xây dựng Hướng dẫn của OECD về Phát triển AI Đáng tin cậy; là thành viên tham gia Tiến trình Hiroshima của G7, công tác chuẩn bị Hội nghị Thượng đỉnh Hành động AI tại Paris, Nhóm Công tác AI Hàn Quốc–Hoa Kỳ, và là thành viên của phái đoàn Hàn Quốc tại Mạng lưới Các Viện An toàn AI Quốc tế.

Lan Xue (Trung Quốc)

Chủ tịch Ủy ban chuyên gia quốc gia về quản trị AI; Viện trưởng Viện Quản trị Quốc tế về AI tại Đại học Thanh Hoa; thành viên Nhóm Tư vấn của Vụ Khoa học, Công nghệ và Đổi mới (STI) của OECD; cố vấn cho Viện An toàn AI Trung Quốc; Đồng Chủ tịch Hội đồng Lãnh đạo của Mạng lưới Giải pháp Phát triển Bền vững Liên Hợp Quốc (UNSDSN); là người nhận Giải thưởng Đóng góp Xuất sắc của Đại học Phục Đán về Khoa học Quản lý và Giải thưởng Đóng góp Xuất sắc từ Hiệp hội Khoa học Khoa học và Chính sách Khoa học – Công nghệ Trung Quốc.

Stuart Russell (Anh và Hoa Kỳ)

Giáo sư xuất sắc ngành Khoa học Máy tính và Giám đốc Trung tâm AI Tương thích với Con người tại Đại học California, Berkeley; tác giả của cuốn *Artificial Intelligence: A Modern Approach* – giáo trình chuẩn về trí tuệ nhân tạo được sử dụng tại 1.500 trường đại học ở 135 quốc gia và được trích dẫn hơn 74.000 lần; Đồng Chủ tịch Nhóm chuyên gia về Tương lai của AI thuộc OECD và Hội đồng Toàn cầu về AI của Diễn đàn Kinh tế Thế giới.

Jaan Tallinn (Estonia)

Thành viên của Nhóm Cố vấn về AI của Liên Hợp Quốc; từng phục vụ trong Nhóm Chuyên gia Cấp cao về AI của Ủy ban Châu Âu; Đồng sáng lập Trung tâm Nghiên cứu Rủi ro Tồn vong tại Đại học Cambridge và Viện Tương lai Cuộc sống (cả hai đều là những tổ chức tiên phong về các vấn đề liên quan đến AGI); Thành viên Hội đồng Quản trị Trung tâm An toàn AI; nhà đầu tư người Estonia trong lĩnh vực an toàn AGI; kỹ sư sáng lập của Skype và FastTrack/Kazaa; đồng thời là giám đốc đầu tư sáng lập của DeepMind thuộc Google.

Mariana Todorova (Bulgaria)

Đại diện của Bulgaria trong Nhóm Liên chính phủ về Các Khung Đạo đức AI của UNESCO; người phát ngôn hàng đầu về AGI trên các phương tiện truyền thông Bulgaria; tác giả và diễn giả được quốc tế công nhận về các khía cạnh đạo đức và công nghệ của AI và AGI; cựu Nghị sĩ Quốc hội và cố vấn cho Tổng thống Cộng hòa Bulgaria.

José Jaime Villalobos (Costa Rica)

Trưởng nhóm Quản trị Đa phương tại Viện Tương lai của Cuộc sống (Future of Life Institute); Nghiên cứu viên cao cấp tại Trung tâm Đổi mới Quản trị Quốc tế; Nghiên cứu viên cộng tác tại Sáng kiến Quản trị AI của Oxford Martin; Nghiên cứu viên cộng tác tại Viện Luật & AI (Institute for Law & AI); Tiến sĩ luật quốc tế; đồng tác giả của nhiều cuốn sách và bài báo hàng đầu về quản trị AI quốc tế.

Dịch bởi: Trang Phạm và Thu Phượng.