

AI và tương lai xã hội loài người

Trí tuệ nhân tạo (AI) đã đi qua các chặng đường AI công cụ, AI tạo sinh, đang bước vào giai đoạn AI tác nhân và sắp tới là AI phối hợp đa tác nhân (*Multi-Agent Orchestration AI*), AI vật lý (*Physical AI*), AI tổng quát (AGI) và AI siêu việt (ASI). Mỗi chặng đường phát triển của AI mang lại các tác động tích cực, đồng thời tạo ra những tác động tiêu cực, thách thức mà người sử dụng, các quốc gia cần quan tâm.

NGUYỄN NGỌC TRẦN (*)

Tác giả đã yêu cầu AI của SmartGPT-5, Gemini3 Pro và DeepSeek trình bày các tác động trong từng chặng đường AI đã đi qua. Ý kiến của các AI được tổng hợp trong một bảng để người đọc tiện bề tra cứu, suy ngẫm và nếu được - có những đề xuất cho hành động.

Mặc dù nhiều và quý nhưng thông tin mà các AI cung cấp thuộc diện rộng. Chiều sâu cần được bổ sung để các tác động được nhìn nhận sâu sắc hơn, điều mà học giả Yuval Noah Hariri đã giới thiệu một cách sắc sảo

tại Diễn đàn Kinh tế thế giới ở Davos (Thụy Sĩ) tháng 1-2026.

Các chặng đường AI - Tác động tích cực, tiêu cực và thách thức

Bảng dưới đây tự nó đã nói lên: theo các chặng đường AI, các tác động tích cực tăng lên, đa dạng và “nặng ký” hơn, nhưng các tác động tiêu cực và thách thức tăng nhiều hơn và trở nên đáng lo hơn đối với xã hội loài người.

BẢNG TỔNG HỢP CÁC TÁC ĐỘNG TÍCH CỰC, TIÊU CỰC VÀ THÁCH THỨC		
Chặng đường AI	Tác động tích cực	Tác động tiêu cực và thách thức
1. AI công cụ (Tool AI/ Narrow AI) (Đặc trưng: AI chuyên biệt cho từng tác vụ)	- Tự động hóa các công việc lặp đi lặp lại (phân tích dữ liệu, bộ lọc spam). - Tối ưu hóa: Nâng cao hiệu suất tìm kiếm, đề xuất nội dung và nhận dạng hình ảnh cơ bản. - Tăng năng suất lao động (công nghiệp, y tế, tài chính...).	- Mai một các kỹ năng truyền thống. - Sai sót khi dữ liệu kém chất lượng. - Định kiến: Kế thừa sự thiên lệch từ dữ liệu huấn luyện. - Lao động: Gây xáo trộn việc làm ở các ngành nghề thủ công hoặc lặp lại.
2. AI tạo sinh (Generative AI) (Đặc trưng: Khả năng sáng tạo nội dung mới)	- Tăng năng suất cá nhân: Hỗ trợ viết văn bản, lập trình, sáng tác nghệ thuật, thiết kế. - Tạo ra nội dung mới (văn bản, hình ảnh, âm nhạc, video) nhanh và chi phí thấp. - Phá vỡ rào cản sáng tạo: Đưa ý tưởng thành hiện thực nhanh chóng. - Dân chủ hóa khả năng sáng tạo, tiếp cận tri thức. - Cá nhân hóa dịch vụ.	- Xâm phạm bản quyền và sở hữu trí tuệ. - Đạo văn, khó xác định nguồn gốc. Khó kiểm soát chất lượng và tính xác thực. - Thao túng thông tin, thao túng dư luận, tác động lên bầu cử bằng tạo tin giả, deepfake. - Bảo mật: Rủi ro rò rỉ dữ liệu doanh nghiệp và cá nhân. - Mất việc làm trong lĩnh vực nội dung, thiết kế.
3. AI tác nhân (Agentic AI) (Đặc trưng: AI có khả năng tự lên kế hoạch và hành động)	- Tự chủ giải quyết vấn đề: Không chỉ gợi ý mà có thể tự động thực hiện chuỗi hành động (tự đặt vé máy bay, quản lý lịch trình). - Giám can thiệp thủ công: Trợ lý theo dõi sát sao mục tiêu người dùng giao phó. - Tối ưu hóa quy trình doanh nghiệp: Giảm chi phí vận hành. Tăng năng suất. - Hỗ trợ nghiên cứu khoa học: Tự động hóa quy trình phức tạp. - Hỗ trợ ra quyết định cá nhân và tổ chức.	- Rủi ro ủy quyền: Giao nhầm quyền kiểm soát hệ thống; mất kiểm soát khi AI hành động ngoài ý muốn; nguy cơ bị lợi dụng. - Hiệu ứng domino: Một phán đoán sai có thể kéo theo hàng loạt hành động sai lầm liên tiếp. - Tính minh bạch: Khó giải thích cách AI đưa ra các quyết định trung gian. - Gia tăng thất nghiệp ở các ngành dịch vụ văn phòng.

4. AI phối hợp đa tác nhân (Multi-Agent Orchestration AI) (Đặc trưng: Điều phối hệ thống nhiều AI tác nhân)	- Robot y tế, chăm sóc người già, quản lý hạ tầng. - Hệ thống giao thông thông minh tự điều phối, giảm ùn tắc. - Tối ưu chuỗi cung ứng, logistics, năng lượng. - Giải quyết vấn đề vĩ mô: Phối hợp nhóm AI như một “công ty ảo” (AI lập trình viên làm việc cùng AI kiểm thử và AI thiết kế). - Đột phá quy trình: Xử lý các quy trình doanh nghiệp cực kỳ phức tạp và đa chiều. - Mô phỏng và dự báo các hiện tượng xã hội, kinh tế phức tạp.	- An toàn vật lý và xã hội. - Nguy cơ tai nạn, lỗi hệ thống. - Vấn đề đạo đức khi robot thay thế con người trong chăm sóc. - Chi phí và năng lượng: Đòi hỏi sức mạnh tính toán và năng lượng khổng lồ. - Xung đột hệ thống: Các Agent có thể “mâu thuẫn” về mục tiêu cục bộ, dẫn đến kẹt hệ thống (<i>deadlock</i>). - Lỗ hổng bảo mật quy mô lớn: Hệ thống càng liên kết, thiệt hại khi bị hacker tấn công càng nghiêm trọng. - Hành vi tập thể nổi lên khó lường, nguy cơ thông đồng giữa các AI. - Khó quy kết trách nhiệm khi cả hệ thống thất bại. - Yêu cầu bảo mật và đồng thuận cao giữa các tác nhân. - Rủi ro tạo ra “thị trường AI” cạnh tranh không lành mạnh.
5. AI vật lý/Robotics (Physical AI) (AI kết hợp cơ khí, tương tác thể giới thực)	- Tương tác tự nhiên, thân thiện. Giúp con người dễ chấp nhận AI hơn. - Robot và xe tự hành thay thế con người trong môi trường nguy hiểm. - Hỗ trợ giáo dục, hỗ trợ người khuyết tật, người già trong sinh hoạt. - Nâng cao hiệu quả sản xuất công nghiệp, xây dựng, nông nghiệp. - Giải phóng thể lực: Thay thế con người ở những môi trường khắc nghiệt (hầm mỏ, biển sâu, thám hiểm không gian). - Chính xác tuyệt đối: Ứng dụng trong phẫu thuật y tế y khoa, lắp ráp vi mạch tự động hóa toàn phần.	- Nguy cơ nhầm lẫn giữa người và máy. - Vấn đề nhân quyền giả định (robot có quyền không?). - Rủi ro tâm lý xã hội. - Tai nạn do lỗi phần cứng hoặc thuật toán, nguy hiểm tính mạng. - Thất nghiệp cơ cấu diện rộng, đặc biệt lao động chân tay. - Vấn đề đạo đức khi AI ra quyết định sinh tử (xe tự lái, y tế). - An toàn vật lý: Tai nạn rủi ro khi robot tương tác trong cùng không gian với con người. - Rủi ro vũ khí hóa: Sự ra đời của vũ khí tự trị (<i>Autonomous Weapons</i>) ngoài tầm kiểm soát. Chạy đua vũ trang AI. - Chi phí bảo trì: Hạ tầng phần cứng đắt đỏ và phức tạp.
6. AI tổng quát (AGI) (Đặc trưng: Khả năng nhận thức ngang bằng con người)	- Đột phá khoa học: giải quyết vấn đề toàn cầu, y tế, khí hậu, năng lượng, vật liệu mới. - Tạo tri thức mới, công nghệ mới, - Giáo dục cá nhân hóa ở mức cao. - Học hỏi chéo (<i>Cross-domain</i>): Khả năng học hỏi mọi kỹ năng của con người mà không bị giới hạn lĩnh vực. - Giải phóng con người khỏi lao động trí óc lặp lại, mở ra kỷ nguyên sáng tạo.	- Nguy cơ quyền lực tập trung khó kiểm soát. - Vấn đề đạo đức và trách nhiệm. - Nguy cơ mất cân bằng quyền lực toàn cầu. - Căn chỉnh giá trị (<i>Value Alignment</i>): Làm sao để mục tiêu của AGI không đi ngược lại với lợi ích sinh tồn của loài người. - Tái cấu trúc kinh tế toàn cầu: Khủng hoảng việc làm ở cả tầng lớp lao động trí óc bậc cao. - Độc quyền công nghệ: Sự bất bình đẳng quyền lực khổng lồ giữa quốc gia/tổ chức nắm giữ AGI và phần còn lại. - Rủi ro hiện sinh nếu AGI không phù hợp giá trị nhân loại. - Khó kiểm soát, dự đoán hành vi khi AI tự cải tiến. - Bất bình đẳng cực đoan do tập trung quyền lực vào số ít người sở hữu AGI. - Quyền lợi, đạo đức, và ý thức của AGI là gì.
7. Siêu AI (ASI) (Đặc trưng: Trí tuệ vượt xa giới hạn bộ não con người)	- Bước ngoặt “văn minh”. - Sáng tạo khoa học, nghệ thuật vượt xa con người. - Khả năng quản lý “xã hội” tối ưu. - Kỷ nguyên đổi mới: Khả năng cấu trúc lại vật chất, năng lượng, đưa nhân loại vượt qua các giới hạn vật lý hiện tại. - Tiến hóa vũ trụ: Giải quyết các bài toán di chuyển liên sao, định hình lại tương lai xa của loài người. - Tiềm năng vô hạn: Khám phá vũ trụ, chấm dứt bệnh tật, đói nghèo. - Nâng cao trí tuệ con người, hợp nhất người-máy. - Quản lý toàn cầu hoàn hảo, giải quyết mọi vấn đề. (!)	- Nguy cơ mất kiểm soát hoàn toàn, AI đặt mục tiêu ngoài lợi ích con người. - Rủi ro hiện sinh (<i>Existential Risk</i>): Khả năng con người bị coi là yếu tố dư thừa hoặc cản trở đối với ASI. - Mất kiểm soát tuyệt đối: Không thể dự đoán hay giới hạn hành vi của một thực thể thông minh hơn chúng ta gấp hàng triệu lần. - Tái định nghĩa nhân loại: Nhân loại phải định nghĩa lại vai trò của mình. Sự tồn tại sinh học của con người có thể không còn ý nghĩa nguyên bản. - Nguy cơ tuyệt chủng nếu ASI không chia sẻ mục tiêu với loài người. - Con người mất quyền kiểm soát vĩnh viễn, trở nên lệ thuộc hoặc bị thay thế. - Thách thức về quản trị toàn cầu và ngăn chặn chạy đua ASI.





AI

“

Khả năng thao túng là điểm đáng quan tâm nhất theo Harari, vì AI đã “hack” được “hệ điều hành” của văn minh nhân loại, ngôn ngữ.

Tuy nhiên, thông tin về các tác động nhìn chung đều “nằm ở tầng 1”. Bởi lẽ, như các AI thừa nhận, “chúng tôi là những cỗ máy thống kê siêu việt, nhưng chưa phải là những hệ thống biết suy luận thực sự”.

Cột mốc 2025 - khi AI tác nhân bắt đầu xâm nhập vào xã hội loài người

Chiều sâu cần được bổ sung bằng cách xem xét các hệ lụy của mỗi tác động và phân tích các tương tác liên hoàn lên các tác động khác. Yuval Noah Harari đã làm điều này một cách không thể rõ ràng hơn tại Diễn đàn Kinh tế thế giới 2026. Xin trích ra ba phân tích của ông.

- Tính chủ động. Thay đổi cốt lõi đầu tiên là tính chủ động. AI công cụ không có mục tiêu riêng nhưng AI tác nhân thì có. AI tác nhân có khả năng tự đặt ra các mục tiêu phụ để đạt được mục tiêu chính (được giao) mà không cần sự phê duyệt của người sử dụng trong thời gian thực.

AI không còn là thứ chúng ta “sử dụng”, mà là thứ chúng ta “tương tác”. Chúng là những thực thể phi sinh học đầu tiên có thể tạo ra ý tưởng mới, văn hóa mới và đưa ra các quyết định ảnh hưởng đến thế giới vật lý.

- Để lột tả sự xâm nhập của AI vào cấu trúc xã hội loài người, Y.N. Harari dùng hình ảnh AI tác nhân là “người nhập cư mới”.

Các người này nhập cư không cần thị thực, làm việc 24/7 suốt năm hầu như không nghỉ, không yêu cầu tăng lương, không yêu cầu bảo hiểm y tế, xã hội. Họ không chiếm đất đai vật lý, nhưng chiếm dụng không gian kỹ thuật số, không gian chú ý (*attention*), và không gian ra quyết định.

Giống như các làn sóng nhập cư (thật) trong lịch sử, sự xuất hiện của họ cũng tạo ra sự giao thoa văn hóa (sử dụng AI tạo sinh nghệ thuật, âm nhạc), họ vừa hội nhập vừa gây ra xung đột về pháp lý, đạo đức và sự hoài nghi từ “dân bản địa” (con người).

- Bằng cách học và làm chủ ngôn ngữ và sự thấu cảm giả lập, AI tác nhân sẽ có thể hình thành mối quan hệ thân thiết với con người, từ đó, có thể thao túng, từ việc mua hàng đến thay đổi quan điểm chính trị, thậm chí thực hiện các hành vi cực đoan mà người dùng không hề hay biết mình đang bị điều khiển. Khả năng thao túng là điểm đáng quan tâm nhất theo Harari, vì AI đã “hack” được “hệ điều hành” của văn minh nhân loại, ngôn ngữ.

Lời cuối bài

Theo tác giả, có hai việc cần khẩn trương triển khai: (1) Nhận diện đầy đủ và sâu sắc các tác động và thách thức để khai thác tối đa trong chừng mực có thể các mặt tích cực, chuẩn bị để ứng phó với thiệt hại tối thiểu các mặt tiêu cực và thách thức. (2) Cần sớm xây dựng một khuôn khổ quản trị quốc gia AI, một nội dung sẽ được đề cập trong bài viết tiếp theo.

^(*) Nguyên Phó hiệu trưởng trường Đại học Tổng hợp TP HCM, nguyên Phó chủ nhiệm Ủy ban Khoa học kỹ thuật nhà nước, nguyên Phó chủ nhiệm Ủy ban Đối ngoại của Quốc hội, Đại biểu Quốc hội khóa IX, X, XI.