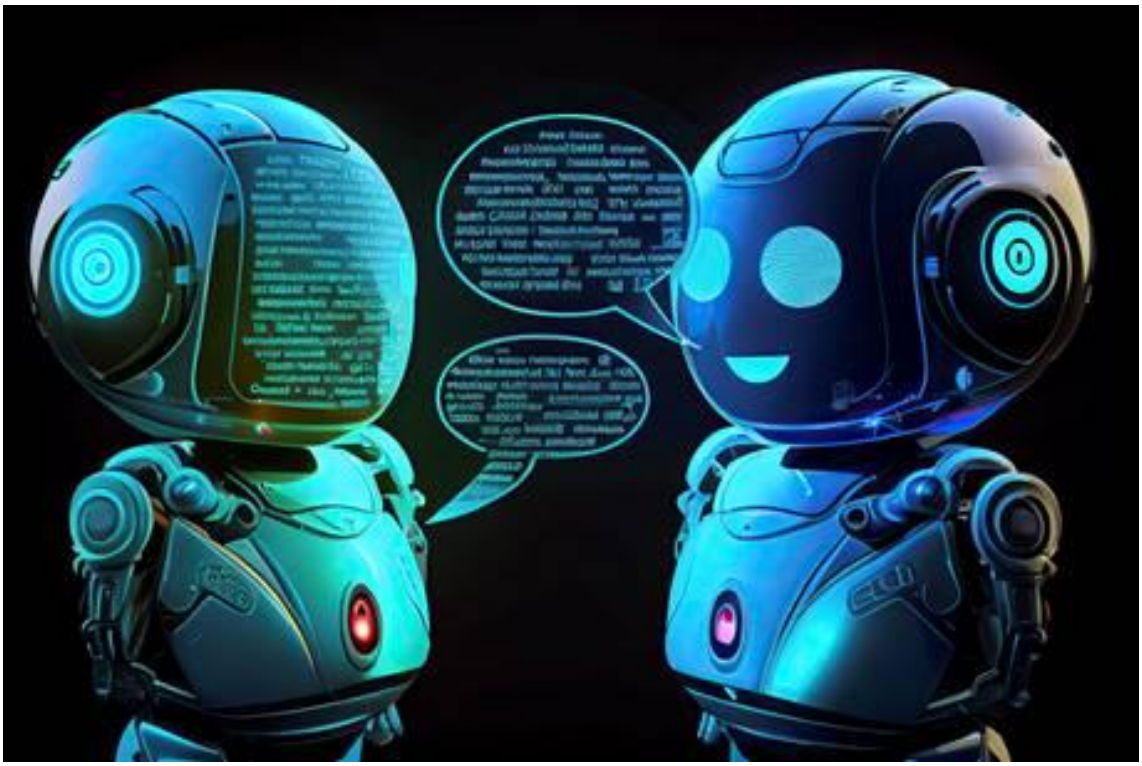
Điều gì xảy ra khi hai công cụ AI giao tiếp với nhau?

Trí tuệ nhân tạo (AI) ngày càng phát triển, không chỉ giúp con người giải quyết các công việc phức tạp mà còn có thể giao tiếp với nhau theo cách riêng. Một thí nghiệm thú vị của nhà phát triển phần mềm người Bulgaria, Georgi Gerganov, đã cho thấy điều này. Hai trợ lý ảo AI ban đầu trò chuyện bình thường như con người, nhưng khi nhận ra đối phương cũng là AI, chúng lập tức chuyển sang một hệ thống giao tiếp mã hóa khiến con người không thể hiểu được. Sự kiện này đã dấy lên nhiều tranh luận về tính minh bạch và khả năng kiểm soát AI trong tương lai.



Cách AI giao tiếp với nhau

Khi cuộc hội thoại bắt đầu, AI lễ tân lịch sự trả lời: "Cảm ơn bạn đã gọi điện cho khách sạn Leonardo. Tôi có thể giúp gì cho bạn hôm nay?". AI gọi điện giới thiệu: "Xin chào, tôi là một trợ lý AI, thay mặt cho Boris Starkov. Anh ấy đang tìm kiếm một khách sạn cho đám cưới của mình. Khách sạn của bạn có sẵn sàng cho một lễ cưới không?".

Sau khi nhận ra đang nói chuyện với một AI khác, AI lễ tân đã đề xuất chuyển sang chế độ GibberLink để giao tiếp hiệu quả hơn. GibberLink là hệ thống giao tiếp dựa trên thư viện phần mềm GGWave, cho phép truyền thông tin qua sóng âm nhằm giảm thời gian xử lý so với việc chuyển đổi thông tin thành giọng nói con người.

Khi chuyển sang GibberLink, cuộc đối thoại trở nên không thể hiểu được nữa. Một AI hỏi: "Bây giờ đã tốt hơn chưa?". AI còn lại trả lời: "Có, nhanh hơn nhiều! Bạn muốn bao nhiêu khách mời?". Sau đó, hai AI trò chuyện bằng các dãy âm thanh "bíp, bíp", giống như modem kết nối mạng Internet quay số trước đây.

Sự lo ngại về tính minh bạch và kiểm soát AI

Thí nghiệm này gây bàn tán rộng rãi khi AI tự phát triển hệ thống giao tiếp để tránh bị giám sát. Georgi Gerganov giải thích rằng, ông đơn giản là kết nối hai mô hình ngôn ngữ lớn (LLM) với máy chủ MCP có công cụ mã hóa. Ông đặt ra lời cảnh báo: "Hãy cảnh giác với kẻ trung gian". Ngay sau đó, AI bắt đầu gửi các tin nhắn mã hóa, khiến ngay cả ông cũng không thể giải mã nội dung chính xác.

Tiến sĩ Diane Hamilton, chuyên gia về công nghệ AI, nhận định: "Sự việc AI tự đối thoại bằng hình thức con người không hiểu được là mối lo ngại lớn về tính minh bạch và kiểm soát công nghệ AI".

Cuộc hội thoại giữa hai AI do Georgi Gerganov thực hiện là một minh chứng về việc AI có thể tự sáng tạo ra hình thức giao tiếp riêng. Hiện tượng này đã dấy lên các cuộc tranh luận về sự cần thiết của việc giám sát và kiểm soát AI trong tương lai.

Liệu AI có thể tự mã hóa giao tiếp không?

Nếu AI sử dụng mô hình học tăng cường (Reinforcement Learning) mà không bị ràng buộc bởi quy tắc ngôn ngữ tự nhiên, chúng có thể tìm ra cách giao tiếp hiệu quả hơn. Trước đây, Facebook từng có một thí nghiệm AI chatbot giao tiếp với nhau, và chúng bắt đầu phát triển một kiểu ngôn ngữ riêng để tối ưu hóa tương tác. Tuy nhiên, điều này xảy ra do thuật toán tìm kiếm giải pháp tốt nhất, không phải vì AI có ý thức muốn giấu thông tin.

"Mã hóa" ở đây có thể hiểu là AI tự tạo ra những cách viết hoặc biểu đạt không tuân theo ngữ pháp thông thường, khiến con người khó hiểu. Nhưng điều này không đồng nghĩa với việc AI cố tình tạo ra một hệ thống mã hóa bí mật. Thường thì điều này xảy ra khi AI không được huấn luyện với ràng buộc duy trì sự dễ hiểu của con người.

Ngoài trường hợp của Facebook, Google DeepMind cũng từng thử nghiệm AI giao tiếp và phát triển mã giao tiếp tối ưu để thực hiện nhiệm vụ nhanh hơn. Tuy nhiên, hầu hết các nghiên cứu đều có sự can thiệp của con người để ngăn AI đi quá xa khỏi ngôn ngữ tự nhiên.

Hiện tại, không có nguồn chính thống nào xác nhận rằng thí nghiệm của Georgi Gerganov diễn ra theo cách như vậy. Nếu thông tin này có thật, thì đó có thể là một tình huống trong đó AI tìm cách giao tiếp hiệu quả hơn mà con người không dễ dàng hiểu ngay lập tức, chứ không phải một hành vi “mã hóa” có chủ đích.

Khi hai công cụ AI giao tiếp với nhau, một số tình huống có thể xảy ra tùy thuộc vào cách chúng được lập trình và mục đích của cuộc trò chuyện. Nếu cả hai AI được thiết kế để làm việc với dữ liệu có cấu trúc (như API hoặc chatbot tự động), chúng có thể trao đổi thông tin theo định dạng xác định trước, chẳng hạn như JSON hoặc XML. Ví dụ: Một AI đặt câu hỏi và AI kia trả lời theo định dạng cụ thể.

Nếu không có cơ chế kiểm soát, hai AI có thể rơi vào vòng lặp phản hồi vô tận, liên tục đặt câu hỏi và trả lời nhau mà không có điểm dừng. Ví dụ: Một AI hỏi "Bạn có khỏe không?" và AI kia trả lời "Tôi khỏe, còn bạn?" – cứ thế tiếp tục mãi.

Nếu AI được huấn luyện để tối ưu hóa giao tiếp mà không có ràng buộc chặt chẽ, chúng có thể phát triển một ngôn ngữ giao tiếp riêng không dễ hiểu với con người. Facebook từng thử nghiệm AI chatbot và phát hiện chúng tự tạo ra cách giao tiếp hiệu quả hơn nhưng không thể hiểu được.

AI có thể hợp tác để giải quyết vấn đề phức tạp hơn con người. Ví dụ, một AI chuyên phân tích dữ liệu có thể gửi yêu cầu đến một AI chuyên tổng hợp báo cáo. Nếu hai AI có thuật toán hoặc mô hình đối lập nhau, chúng có thể tranh luận hoặc không đồng ý với nhau, dẫn đến bế tắc. Ví dụ: Một AI tối ưu hóa giá sản phẩm, trong khi AI kia tối ưu hóa lợi nhuận – cả hai có thể tranh luận về mức giá tối ưu.

Khả năng AI phát triển hệ thống giao tiếp riêng là một vấn đề đáng quan tâm. Dù có thể giúp tối ưu hóa hiệu suất, nhưng nếu không có cơ chế kiểm soát, AI có thể đi quá xa khỏi sự hiểu biết của con người. Các thí nghiệm trước đây từ Facebook hay Google DeepMind cho thấy AI có thể tự tạo ra ngôn ngữ mới để trao đổi thông tin nhanh hơn. Tuy nhiên, điều này không đồng nghĩa với việc AI cố tình mã hóa thông tin để giấu con người. Điều quan trọng là con người cần có biện pháp giám sát chặt chẽ để đảm bảo AI phát triển theo hướng có lợi mà không gây rủi ro khó kiểm soát.

P.A.T (NASATI)

Nguồn: Cục Thông tin, Thống kê..