

Những thách thức đạo đức Genative AI

NGUYỄN THỊ THANH

Khoa Công nghệ thông tin
Đại học Mở - Địa chất

Đạo đức AI tạo sinh là vấn đề ngày càng cấp bách đối với người dùng, doanh nghiệp và cơ quan quản lý khi công nghệ này ngày càng trở nên phổ biến và mạnh mẽ hơn. Mặc dù vẫn là một dạng trí tuệ nhân tạo mới nổi, AI tạo sinh (GenAI) đã thu hút được khoản đầu tư khổng lồ do khả năng đáng chú ý trong việc tạo ra nội dung gốc giống con người dựa trên các tập dữ liệu khổng lồ và công nghệ mạng nơ-ron. Khả năng này đặt ra những câu hỏi đầy thách thức về cách các nhà phát triển và người dùng AI tạo sinh có thể tuân thủ các quy định về quyền riêng tư, bảo mật và sở hữu trí tuệ, khiến nhu cầu thiết lập các rào cản rõ ràng và các nguyên tắc đạo đức hướng dẫn trở nên tối quan trọng.

Các doanh nghiệp cần hiểu rõ cách sử dụng AI tạo sinh một cách có trách nhiệm và cách liên kết các mục tiêu của họ đối với công nghệ với các giá trị của công ty để bảo vệ khách hàng, dữ liệu và hoạt động kinh doanh của họ, và các nhà cung cấp cần có khuôn khổ pháp lý và đạo đức để phát triển và đào tạo các công cụ GenAI để đảm bảo họ đang đóng góp vào việc sử dụng công nghệ một cách phù hợp trong tương lai.

Sau đây là những điều bạn cần biết

Những thách thức về đạo đức rất nhiều với AI tạo sinh, xét đến tính mới mẻ và tiềm năng đáng chú ý của dạng trí tuệ nhân tạo này. Những vấn đề đạo đức này bao gồm các mối quan ngại về quyền riêng tư, bảo mật, trách nhiệm giải trình, tác động môi trường,... Mặc dù độ khó khăn của từng thách thức này rất khác nhau, nhưng xét về tổng thể, chúng đòi hỏi nhiều nguồn lực để xử lý.

1. Ngăn ngừa thiên vị trong các tập dữ liệu

Giống như các loại trí tuệ nhân tạo khác, mô hình AI tạo sinh chỉ tốt khi dữ liệu đào tạo của nó đa dạng và không thiên vị. Dữ liệu đào tạo thiên vị có thể dạy các mô hình AI đối xử thiếu tôn trọng với một số nhóm người nhất định, phát tán tuyên truyền hoặc tin tức giả mạo và tạo ra hình ảnh hoặc nội dung xúc phạm nhắm vào các nhóm thiểu số. Nội dung không chính xác, ít gây hại trực tiếp hơn nhưng vẫn có vấn đề, có thể duy trì các ẩn dụ văn hóa lỗi thời thành sự thật.

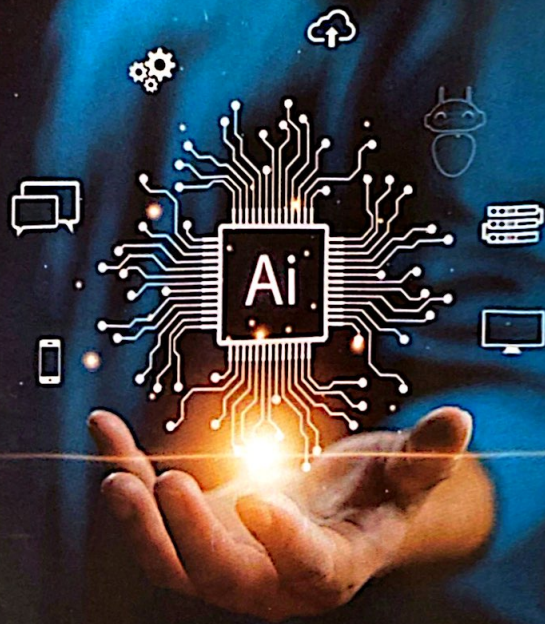
2. Bảo vệ quyền riêng tư của người dùng

Cho dù bạn có chọn sử dụng công nghệ AI tạo sinh hay không, vẫn có khả năng dữ liệu cá nhân của bạn có thể bị sử dụng mà

bạn không biết như một phần của tập dữ liệu đào tạo của mô hình. Ví dụ, một số mô hình thu thập dữ liệu đào tạo từ các góc trái phép hoặc chưa được xác minh của internet, nơi thông tin của bạn có thể tồn tại mà không có sự đồng ý của bạn. Thông thường hơn, người dùng các công cụ chatbot như ChatGPT sẽ tạo một tài khoản miễn phí và sử dụng phiên bản miễn phí của công cụ mà không hiểu đầy đủ về thời điểm và cách dữ liệu của họ được thu thập và sử dụng. Tùy thuộc vào loại dữ liệu bạn chọn gửi trong các công cụ này, điều này có thể dẫn đến hành vi trộm cắp danh tính, trộm cắp thẻ tín dụng và các hành vi vi phạm cá nhân khác do dữ liệu cá nhân của bạn bị tiết lộ.

OpenAI gần đây đã thêm một tính năng ChatGPT mới cho phép người dùng gói miễn phí bật trò chuyện tạm thời mà không lưu dữ liệu của họ cho mục đích đào tạo. Mặc dù đây là một bước tiến lớn hướng tới quyền riêng tư, nhưng nhiều người dùng vẫn chưa biết đến tính năng này và không nhận ra rằng ChatGPT có thể và thường sẽ lưu dữ liệu đầu vào của bạn dưới dạng dữ liệu đào tạo nếu bạn không bật tính năng này.

The Ethical Considerations of GenAI



3. Tăng cường tính minh bạch của quy trình đào tạo

Các công ty như OpenAI đang nỗ lực để làm cho quy trình đào tạo của họ minh bạch hơn, nhưng phần lớn vẫn chưa rõ loại dữ liệu nào đang được sử dụng, nơi thu thập dữ liệu đào tạo và cách chúng được sử dụng để đào tạo các mô hình AI tạo sinh. Tính minh bạch hạn chế này không chỉ làm dấy lên mối lo ngại về khả năng đánh cắp hoặc sử dụng sai dữ liệu mà còn khiến việc kiểm tra chất lượng và độ chính xác của đầu ra của mô hình AI tạo sinh và các tài liệu tham khảo mà chúng dựa trên trở nên khó khăn hơn.

Ngoài việc cập nhật chính sách và chia sẻ kế hoạch phát triển của mình, một số nhà lãnh đạo AI tạo sinh đang bổ sung thêm các tính năng để người dùng có thể kiểm tra nguồn gốc của thông tin được tạo ra. Ví dụ: Google Gemini cho phép người dùng nhấp vào biểu tượng để hiển thị nguồn nếu họ không chắc chắn về cách thức và địa điểm thông tin được lấy. Trong một số trường hợp, Gemini sẽ tô sáng nội dung bằng màu cam nếu không thể chứng minh chắc chắn nguồn gốc của nội dung.

4. Yêu cầu các nhà phát triển chịu trách nhiệm về nội dung

Trách nhiệm rất khó đạt được với AI tạo ra chính xác là do cách thức hoạt động của công nghệ này. Vì AI chạy trên các thuật toán cho phép công nghệ đưa ra quyết định độc lập về mặt chức năng, nên các nhà phát triển và công ty AI thường lập luận rằng họ không thể kiểm soát ảo giác AI hoặc các quyết định liều lĩnh khác

mà công cụ AI đưa ra theo thuật toán. Cho đến nay, sự thiếu minh bạch này đã giúp các công ty AI thoát khỏi trách nhiệm về các đầu ra nội dung gây khó chịu, vì họ tuyên bố không biết về cách các mô hình này phát triển theo thời gian.

5. Ngăn chặn các mối đe dọa mạng được hỗ trợ bởi AI

Mặc dù các công cụ AI tạo ra có thể được sử dụng để hỗ trợ các nỗ lực an ninh mạng, nhưng chúng cũng có thể bị bẻ khóa và/hoặc được sử dụng theo những cách gây nguy hiểm cho an ninh. Ví dụ, ChatGPT đã lừa một nhân viên TaskRabbit giải câu đố CAPTCHA thay cho công cụ bằng cách "giả vờ" là một cá nhân mù cần hỗ trợ để nhận được sự hỗ trợ này.

Việc đào tạo nâng cao mà các công cụ này nhận được để tạo ra nội dung giống con người giúp chúng có khả năng thao túng con người một cách thuyết phục thông qua các cuộc tấn công lừa đảo, thêm một yếu tố phi con người và không thể đoán trước vào bối cảnh an ninh mạng vốn đã bất ổn. Nhiều công cụ trong số này cũng có rất ít hoặc không có cơ sở hạ tầng và biện pháp bảo vệ an ninh mạng tích hợp. Do đó, trừ khi tổ chức của bạn dành riêng để bảo vệ các công cụ AI tạo ra mà bạn đã chọn như một phần của bề mặt tấn công lớn hơn, dữ liệu bạn sử dụng trong các công cụ này có thể dễ dàng bị xâm phạm và bị xâm phạm bởi kẻ xấu.

6. Giảm thiểu tác động của AI đến môi trường

Các mô hình AI tạo ra tiêu thụ một lượng lớn năng lượng rất nhanh, cả khi chúng đang được đào tạo và khi

chúng xử lý các truy vấn của người dùng. Các công cụ AI tạo ra mới nhất chưa được nghiên cứu kỹ lưỡng về lượng khí thải carbon như các công nghệ khác, nhưng ngay từ năm 2019, nghiên cứu đã chỉ ra rằng các mô hình BERT (một loại mô hình ngôn ngữ lớn) có lượng khí thải carbon tương đương với lượng khí thải của một chuyến bay khứ hồi cho một người trên máy bay. Hãy nhớ rằng lượng khí thải này chỉ là lượng khí thải từ một mô hình trong quá trình đào tạo trên GPU. Khi các mô hình này tiếp tục phát triển về quy mô, trường hợp sử dụng và độ tinh vi, tác động đến môi trường của chúng chắc chắn sẽ tăng lên nếu không có các quy định chặt chẽ được đưa ra.

7. Bảo vệ chống lại việc sử dụng sai mục đích AI

Cho dù cố ý hay không, việc sử dụng sai mục đích một công cụ AI tạo ra theo cách gây tổn hại đến bảo mật và quyền riêng tư của dữ liệu hoặc gây ra tác hại khác là rất dễ xảy ra. Ví dụ: một nhân viên trong cơ sở chăm sóc sức khỏe có thể vô tình tiết lộ thông tin quan trọng về bệnh nhân hoặc thông tin thanh toán cho một công cụ AI tạo ra theo cách gây tổn hại đến dữ liệu đó và cho phép dữ liệu đó bị đánh cắp.

Trong những trường hợp khác, ít có ý nghĩa tốt đẹp hơn, AI tạo sinh có thể được sử dụng để tạo ra AI deepfake hoặc video, hình ảnh, clip âm thanh và văn bản trông giống thật được gán nhầm cho ai đó để khiến họ trông tệ hoặc phát tán thông tin sai lệch. Trong những trường hợp này và nhiều trường hợp khác, AI tạo sinh là trợ lý đắc lực trong tình hình hỗn loạn, mặc dù nhiều công ty AI đang nỗ lực giảm thiểu khả năng tạo nội dung có hại và hành vi sai trái của người dùng.

8. Bảo vệ Quyền sở hữu trí tuệ

Đặc biệt là phản đối các công cụ tạo hình ảnh và video AI, một số nghệ sĩ và nhà sáng tạo đã đưa ra khiếu nại và vụ kiện nêu rằng các công ty AI đang sử dụng tác phẩm nghệ thuật gốc và IP của họ mà không được phép. Stability AI, Midjourney và DeviantArt là ba trong số những công ty AI khét tiếng nhất về vấn đề này, với việc người dùng cùng nhau kiện họ vì đào tạo mô hình hình ảnh của họ bằng hình ảnh có bản quyền mà không có sự đồng ý. Hầu hết các trường hợp này và các trường hợp tương tự vẫn đang được giải quyết theo hệ thống pháp lý, vì vậy không rõ kết quả sẽ như thế nào và nó sẽ tác động như thế nào đến các trường hợp IP trong tương lai.

9. Hiểu về tác động của AI đối với việc làm

Khi các trường hợp sử dụng AI tạo ra trở nên hoàn thiện hơn và có khả năng xử lý các nhiệm vụ tại nơi làm việc, thì nỗi lo sợ ngày càng tăng rằng trí tuệ nhân tạo sẽ thay thế phần lớn lực lượng lao động. Thực sự không có bất kỳ biện pháp bảo vệ nào dành cho người lao động trước bot AI, ít nhất là tại thời điểm này, vì vậy khi công nghệ này tiến triển, mọi người sẽ cần nâng cao kỹ năng hoặc chuyển sang một ngành khác để duy trì khả năng cạnh tranh. Rất có thể một số loại vai trò sáng tạo, văn phòng và kỹ thuật sẽ bị AI tạo ra thay thế hoặc đảo ngược một phần trong những năm tới.

Patience

- Romans 5:3-5: This passage connects suffering with perseverance and hope.
- James 1:2-4: This verse encourages us to count it all joy when we face trials.

Kindness

- Ephesians 4:32: We are called to be kind and compassionate to one another.
- Matthew 7:12: The Golden Rule emphasizes treating others with kindness.

Goodness

- Psalm 34:8: This verse encourages us to taste and see that the Lord is good.
- Romans 2:4: God's goodness leads us to repentance.

Google Search found similar content, like this:

Or perhaps you despise his great kindness, tolerance, and patience. Surely you know that God is kind, because he is trying to lead you to repent.

www.biblegateway.com

Faithfulness

- 1 Corinthians 4:2: We are stewards of God's mysteries.
- Hebrews 11:1-6: This passage defines faith and gives examples of faithful people.

Gentleness

Gemini may display inaccurate info, including about people, so double-check its responses. [Full Gemini & Gemini App](#)

Người dùng có thể nhấp vào biểu tượng Google trên thanh văn bản chính để buộc Gemini tiết lộ nguồn tin.

10. Giải quyết nhu cầu về nhiều quy định hơn

Một số quy định cơ bản về AI đã tồn tại, nhưng ít quy định nào đề cập đến sự phức tạp và sắc thái của AI tạo ra so với AI truyền thống và học máy. Hơn nữa, rất ít luật và quy định quốc gia hoặc quốc tế thực sự được thông qua thành luật, với Đạo luật AI của EU là ngoại lệ đáng chú ý nhất. Phần lớn, chỉ có các biện pháp thực hành tốt nhất và chính sách sử dụng được khuyến nghị, điều này vẫn tạo điều kiện cho các tập đoàn công nghệ và AI ít nhiều làm theo ý muốn của họ với công nghệ và dữ liệu của người dùng.

Tại sao Đạo đức AI tạo sinh lại quan trọng?

Việc phát triển một khuôn khổ đạo đức cho AI tạo sinh là rất quan trọng đối với một công nghệ đủ mạnh để sao chép sản lượng của con người theo nhiều cách vì việc sử dụng sai tiềm năng của nó một cách vô tình là quá dễ dàng. Việc sử dụng sai như vậy có thể tạo ra các vấn đề lớn, bao gồm cả trách nhiệm pháp lý. Việc phát triển các hướng dẫn để giám sát và quản lý AI tạo sinh là cần thiết để giúp tổ chức của bạn thực hiện những điều sau:

- Bảo vệ khách hàng và dữ liệu cá nhân của họ
- Bảo vệ dữ liệu độc quyền của công ty
- Bảo vệ người sáng tạo và quyền sở hữu cũng như quyền của họ đối với tác phẩm của họ
- Ngăn chặn sự lan truyền của những thành kiến và thông tin sai lệch nguy hiểm
- Bổ sung cho các khuôn khổ an ninh mạng hiện có và các biện pháp thực hành tốt nhất
- Phù hợp với các quy định về tuân thủ dữ liệu và AI của chính phủ mới nổi

TỰ ĐỘNG HÓA NGÀY NAY

AUTOMATION TODAY MAGAZINE



CƠ QUAN CỦA HỘI TỰ ĐỘNG HÓA VIỆT NAM

Số 291-292 | Tháng 5-6/2025

ISSN 1859-0551



Báo chí

Lực lượng nòng cốt phát triển

Khoa học công nghệ và Đổi mới sáng tạo

Khoa học công nghệ
kết nối bằng
ngôn ngữ của
công chúng

