Ngũ giới và giữ giới trong kỷ nguyên AI

ISSN: 2734-9195 08:30 18/04/2026

Thế giới hiện nay đang bước vào giai đoạn công nghệ phát triển nhanh hơn khả năng thích ứng về đạo đức. AI không chỉ thay đổi cách chúng ta làm việc, mà còn âm thầm định hình lại cách chúng ta suy nghĩ, tương tác và nhận thức về thực tại.

AI đang phát triển mạnh mẽ, tác động sâu rộng đến đời sống xã hội, các vấn đề đạo đức công nghệ ngày càng trở nên cấp thiết.

Ngũ giới - nền tảng đạo đức căn bản của người phật tử, bài viết góp phần gợi mở một hệ quy chiếu nhân văn cho việc phát triển và ứng dụng AI trong thời đại số.

Trí tuệ nhân tạo đặt lại câu hỏi về đạo đức

Sự phát triển nhanh chóng của AI tạo ra bước tiến vượt bậc trong khoa học - công nghệ, đặt ra những vấn đề liên quan đến đạo đức, bao gồm: tính minh bạch (transparency), tính công bằng (fairness), trách nhiệm giải trình (accountability), quyền riêng tư (privacy) và nguyên tắc không gây hại (non-maleficence).

Trước những thực tại đó, việc tiếp cận các truyền thống đạo đức, đặc biệt là Phật giáo, có thể mang lại góc nhìn nhân bản.

Ngũ giới - năm nguyên tắc đạo đức căn bản của Phật giáo có thể được xem như hệ quy chiếu để nhận diện những giới hạn đạo đức (ethical boundaries) mà AI không nên vượt qua.

Ngũ giới: Từ nền tảng kinh điển đến giá trị phổ quát

Ngũ giới (pañca-sīla) không phải là hệ thống đạo đức được thiết lập mang tính áp đặt, mà được hình thành từ chính bối cảnh xã hội thời Đức Phật Thích Ca Mâu Ni còn tại thế. Trong nhiều bản kinh thuộc tạng Pāli như Kinh Tăng Chi Bộ

(Aṅguttara Nikāya), đức Phật đã chế định các giới luật dựa trên những vấn đề cụ thể phát sinh trong đời sống tăng đoàn và xã hội đương thời.

Ban đầu, giới luật không được đặt ra một cách hệ thống, mà được hình thành dần theo nguyên tắc: "*Do sự kiện mà chế giới*" (tùy phạm tùy chế).

Ngũ giới dành cho cư sĩ tại gia, vì vậy mang tính căn bản, nhằm:

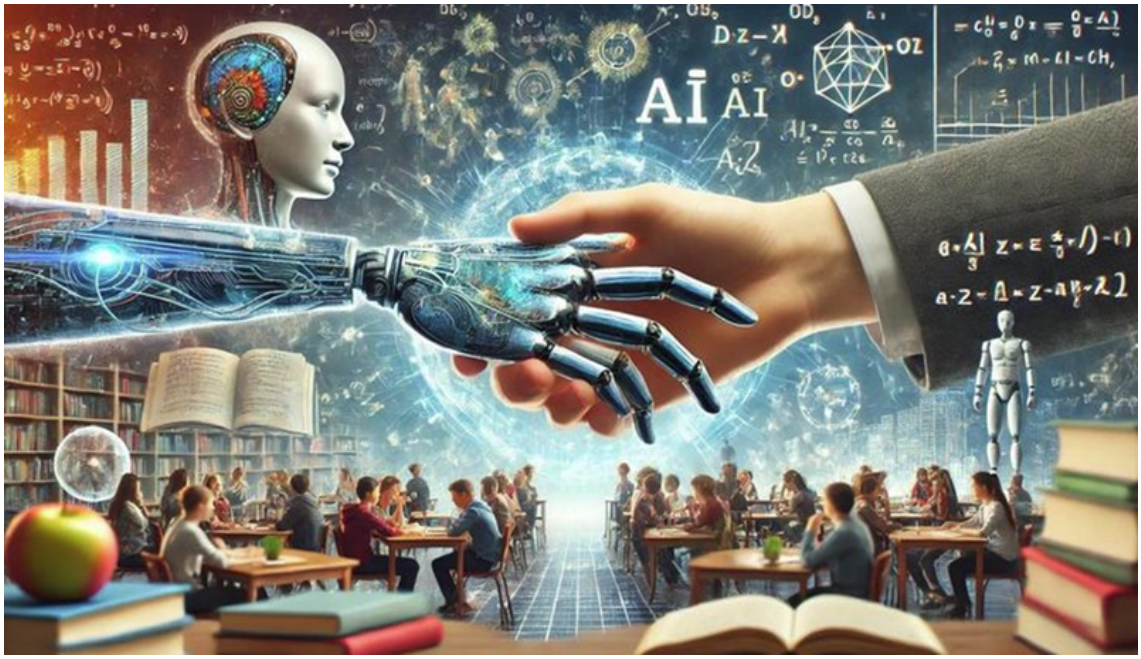
+ Bảo vệ sự sống.
+ Bảo vệ tài sản.
+ Bảo vệ các mối quan hệ.
+ Bảo vệ sự thật.
+ Bảo vệ tâm thức tỉnh táo.

Trải qua hơn 2.500 năm, Ngũ giới không chỉ tồn tại như một quy phạm tôn giáo, mà đã trở thành hệ chuẩn đạo đức có giá trị phổ quát, được nhiều học giả nhìn nhận như phiên dạng đạo đức nhân loại (universal ethics).

Điểm đặc biệt của Ngũ giới là nhấn mạnh đến ý định (cetanā), yếu tố quyết định nghiệp. Điều này khiến Ngũ giới không chỉ điều chỉnh hành vi bên ngoài, mà còn hướng đến sự chuyển hóa nội tâm.

Khác với nhiều hệ thống đạo đức mang tính luật lệ, Ngũ giới nhấn mạnh đến chuyển hóa nội tâm. Một hành vi được xem là thiện hay bất thiện không chỉ dựa trên kết quả, mà còn phụ thuộc vào động cơ và nhận thức của chủ thể hành động. Ngũ giới có thể được hiểu như một *"hệ điều hành đạo đức"* giúp con người duy trì sự hài hòa giữa cá nhân và xã hội.

Khi đặt trong bối cảnh công nghệ, Ngũ giới không trực tiếp áp dụng cho máy móc, nhưng lại có thể đóng vai trò như một tiêu chuẩn để đánh giá cách con người thiết kế, vận hành và sử dụng AI.



Hình ảnh minh hoạ. Ảnh sưu tầm

Ngũ giới và các giới hạn đạo đức của AI

Không sát sinh và giới hạn gây hại của AI

Gần đây, thế giới chứng kiến nhiều cuộc xung đột vũ trang với mức độ công nghệ hóa ngày càng cao. Các chiến dịch quân sự tại Trung Đông, bao gồm những căng thẳng liên quan giữa Mỹ và Iran, hay các xung đột tại Ukraine và Gaza Strip, cho thấy thực tế: công nghệ hiện đại, trong đó có AI, đang được tích hợp sâu vào chiến tranh.

Các hệ thống máy bay không người lái (drone) sử dụng thuật toán để nhận diện mục tiêu, các hệ thống phân tích dữ liệu chiến trường theo thời gian thực, hay thậm chí các nghiên cứu về vũ khí tự hành (autonomous weapons): tất cả đang từng bước làm *“tự động hóa”* hành vi sát hại.

Nhưng điều đáng nói là: dù công nghệ ngày càng chính xác, sai số đạo đức vẫn không thể bị loại bỏ.

Các báo cáo quốc tế nhiều lần ghi nhận việc không kích nhầm mục tiêu dân sự, gây thương vong cho người vô tội. Khi một thuật toán nhận diện sai, hậu quả không phải là *“lỗi hệ thống”*, mà là sinh mạng con người.

Ở đây, vấn đề không còn là kỹ thuật, mà là câu hỏi: Ai chịu trách nhiệm khi một cỗ máy quyết định tước đoạt sự sống?

Giới thứ nhất không chỉ ngăn cấm hành vi sát sinh, mà sâu xa hơn là nuôi dưỡng tâm từ và ý thức bảo hộ sự sống. Khi con người phát triển những công

nghệ có khả năng giết hại từ xa, thậm chí không cần đối diện trực tiếp với nạn nhân, khoảng cách giữa hành động và hệ quả bị kéo giãn và cùng với đó là sự suy giảm của lòng trắc ẩn.

Điều đáng lo không phải là máy móc trở nên nguy hiểm, mà là con người dần trở nên vô cảm trước sự sống, khi việc sát hại được *"trung gian hóa"* qua công nghệ.

Không trộm cắp và quyền sở hữu dữ liệu trong thời đại AI

Trong kỷ nguyên số, dữ liệu cá nhân đã trở thành một loại *"tài sản vô hình"* có giá trị đặc biệt. Tuy nhiên, thực tế cho thấy việc thu thập, sử dụng và khai thác dữ liệu đang diễn ra với tốc độ vượt xa khả năng kiểm soát của con người.

Một trong những vụ việc điển hình là bê bối Cambridge Analytica scandal [1], khi hàng chục triệu dữ liệu người dùng bị khai thác để phục vụ mục đích chính trị mà không có sự đồng thuận rõ ràng. Đây không chỉ là một sự cố kỹ thuật, mà là một biểu hiện rõ nét của việc chiếm đoạt thông tin cá nhân trong không gian số.

Gần đây hơn, nhiều hãng công nghệ lớn bị kiện vì sử dụng dữ liệu từ internet (bao gồm cả sách, bài viết, hình ảnh) để huấn luyện AI mà không xin phép tác giả. Điều này đặt ra câu hỏi nghiêm trọng về quyền sở hữu trí tuệ (intellectual property) và ranh giới giữa "*học hỏi*" và "*chiếm dụng*".

AI "*lấy dữ liệu*" không còn là hành vi cụ thể như trộm cắp truyền thống, mà trở thành một quá trình âm thầm, hệ thống và khó nhận diện.

Giới thứ hai không chỉ dừng lại ở việc không lấy của không cho, mà còn hàm chứa tinh thần tôn trọng quyền sở hữu và sự chính đáng của người khác. Khi dữ liệu cá nhân bị thu thập và khai thác mà không có sự hiểu biết và đồng thuận đầy đủ, con người bị tước đi quyền kiểm soát chính "*bản thân số*" của mình.

Đây không chỉ là vấn đề pháp lý, mà là sự suy giảm của chính mạng (right livelihood) trong môi trường số khi lợi ích kinh tế được đặt lên trên sự tôn trọng con người.

Không tà dâm và sự lệch chuẩn của các mối quan hệ trong không gian số

Sự phát triển của AI đã mở ra khả năng tái tạo hình ảnh và con người với độ chân thực gần như tuyệt đối. Công nghệ deepfake (tạm dịch: công nghệ giả mạo hình ảnh/video bằng AI) đã được sử dụng để tạo ra các video giả mạo, trong đó nhiều trường hợp liên quan đến việc ghép khuôn mặt của người thật vào nội dung nhạy cảm.

Một ví dụ gây chấn động là vụ lan truyền hình ảnh deepfake khiêu dâm nhắm vào Taylor Swift (2024) [2], thu hút hàng triệu lượt xem trước khi bị gỡ bỏ, cho thấy mức độ tổn hại nghiêm trọng đến danh dự và tâm lý mà công nghệ này có thể gây ra. Theo nhiều nghiên cứu quốc tế, khoảng 90% nội dung deepfake trên internet liên quan đến khiêu dâm không đồng thuận (non-consensual pornography - tạm dịch: nội dung khiêu dâm không có sự đồng thuận).

Bên cạnh đó, sự xuất hiện của các chatbot "*bạn đồng hành*" (AI companion - tạm dịch: AI đồng hành) cũng đang làm thay đổi cách con người thiết lập mối quan hệ. Một số người dần lệ thuộc vào những tương tác "*ảo*", nơi mọi phản hồi đều được thiết kế để làm hài lòng, dẫn đến hiện tượng gắn bó cảm xúc với AI (emotional attachment to AI - tạm dịch: gắn kết cảm xúc với trí tuệ nhân tạo).

Điều này đặt ra câu hỏi sâu sắc: Khi mối quan hệ không còn dựa trên sự thật và tương tác thật, thì giá trị mối quan hệ đó là gì?

Luận chứng này cho thấy, Giới thứ ba không chỉ liên quan đến hành vi tình dục, mà còn bảo vệ sự chân thật và lành mạnh của các mối quan hệ. Khi công nghệ tạo ra những *"thực tại giả"*, con người dễ rơi vào trạng thái mê lầm, không còn phân biệt rõ giữa thật và giả, giữa cảm xúc chân thực và cảm xúc được lập trình.

Đây là biểu hiện của ái dục (tanhā) được công nghệ hóa, một dạng chấp thủ tinh vi nhưng mạnh mẽ, khiến con người xa rời thực tại.



Hình ảnh chỉ mang tính chất minh hoạ. Ảnh sưu tầm

Không nói dối và khủng hoảng sự thật trong thời đại AI

Chưa bao giờ trong lịch sử, con người lại đối diện với *"khủng hoảng sự thật"* sâu sắc như hiện nay. AI có thể tạo ra văn bản, hình ảnh, video với độ chân thực cao, khiến ranh giới giữa thật và giả trở nên mong manh.

Hiện tượng thông tin sai lệch (misinformation - tạm dịch: thông tin sai không chủ đích) và thông tin sai lệch có chủ đích (disinformation - tạm dịch: thông tin sai có chủ ý) ngày càng phổ biến, đặc biệt trong các kỳ bầu cử hoặc khủng hoảng xã hội. Các video deepfake giả mạo chính trị gia hay người nổi tiếng đã từng gây hoang mang dư luận.

Một trường hợp điển hình năm 2025: hai luật sư tại Mỹ đã sử dụng AI để trích dẫn án lệ trong hồ sơ pháp lý, nhưng các án lệ này hoàn toàn không tồn tại, hệ quả của hiện tượng ảo giác AI (AI hallucination - tạm dịch: hiện tượng AI tạo ra thông tin sai nhưng trình bày như thật). Vụ việc buộc tòa án phải xử phạt và trở thành cảnh báo toàn cầu về độ tin cậy của AI.

Điều này nguy hiểm ở chỗ: Người dùng dễ tin và khó kiểm chứng, lời nói chân thật không chỉ là không nói sai, mà còn là nói lời có lợi ích, đúng thời và đúng chỗ. Khi AI khuếch đại thông tin sai lệch, đó không chỉ làm sai lệch nhận thức cá nhân, mà còn làm xói mòn nền tảng niềm tin xã hội.

Từ góc nhìn sâu hơn, *"ảo giác AI"* có thể được xem như một dạng vọng tưởng kỹ thuật số, phản ánh chính sự bất toàn trong nhận thức của con người, khi giao phó việc tìm kiếm chân lý cho máy móc.

Không sử dụng chất gây nghiện và sự lệ thuộc vào công nghệ

Một trong những tác động âm thầm nhưng sâu sắc nhất của AI là khả năng tạo ra sự lệ thuộc hành vi. Các nền tảng mạng xã hội sử dụng thuật toán để cá nhân hóa nội dung, khiến người dùng liên tục bị cuốn vào dòng thông tin bất tận.

Hiện tượng *"cuộn vô hạn"* (infinite scroll - tạm dịch: cuộn nội dung không giới hạn), video ngắn, hay thông báo liên tục đều được thiết kế nhằm kích thích hệ thống phần thưởng dopamine trong não bộ. Điều này liên quan đến cơ chế vòng lặp dopamine (dopamine loop - tạm dịch: chu trình kích thích khoái cảm thần kinh), khiến người dùng quay lại sử dụng nền tảng nhiều lần.

Thực tế, các nền tảng như TikTok, Instagram hay Facebook đã nhiều lần bị các cơ quan lập pháp Mỹ và châu Âu chất vấn về tác động gây nghiện đối với thanh thiếu niên. Nhiều nghiên cứu cũng chỉ ra mối liên hệ giữa việc sử dụng mạng xã hội quá mức và các vấn đề như lo âu, trầm cảm và suy giảm khả năng tập trung.

Điều này dẫn đến: Giảm khả năng tập trung. Gia tăng lo âu. Và hình thành thói quen lệ thuộc.

Giới thứ năm hướng đến việc bảo vệ tâm thức tỉnh táo và tự chủ. Khi con người bị cuốn vào các cơ chế gây nghiện do AI thiết kế, họ dần đánh mất khả năng nhận biết và làm chủ chính mình.

Đây là biểu hiện của vô minh (ignorance - tạm dịch: sự không thấy biết đúng như thật) được nuôi dưỡng bởi công nghệ, khi con người không còn thấy rõ mình đang bị chi phối.



Hình ảnh chỉ mang tính chất minh hoạ. Ảnh sưu tầm

AI không phải là chủ thể đạo đức: Phân tích từ triết học đến Phật học

Một trong những ngộ nhận phổ biến hiện nay là xu hướng *"nhân cách hóa"* AI, xem AI như một thực thể có khả năng suy nghĩ, lựa chọn và chịu trách nhiệm như con người. Tuy nhiên, xét từ cả triết học phương Tây lẫn Phật học, góc nhìn này là không chính xác.

Trong triết học đạo đức hiện đại, một thực thể chỉ được xem là chủ thể đạo đức (moral agent) khi hội đủ các điều kiện:

+ Có ý thức về hành vi của mình.
+ Có khả năng phân biệt đúng - sai.
+ Có năng lực lựa chọn dựa trên chuẩn mực đạo đức.
+ Và có thể chịu trách nhiệm về hậu quả.

AI, dù phức tạp đến đâu, vẫn chỉ là một hệ thống xử lý dữ liệu dựa trên xác suất. AI không có ý định (intention), mà chỉ có tối ưu hóa (optimization). AI không *"muốn"* làm điều thiện hay ác, mà chỉ thực hiện theo mục tiêu được lập trình.

Từ góc nhìn Phật học, yếu tố quyết định đạo đức của một hành vi chính là tâm ý (cetanā). Đức Phật dạy: "*Chính tâm ý là nghiệp*" (cetanāhaṃ bhikkhave kammaṃ vadāmi).

Điều này có nghĩa, một hành vi chỉ mang tính đạo đức khi hành vi đó phát sinh từ ý định có ý thức.

Do đó, AI không thể tạo nghiệp, vì AI không có tâm. Nhưng điều này không làm giảm đi tính nghiêm trọng của các hệ quả mà AI gây ra, ngược lại, làm nổi bật trách nhiệm của con người.

Trong đời sống thực tế, chúng ta thấy rõ điều này:

+ Một thuật toán đề xuất nội dung cực đoan không *"có ác ý"*, nhưng được thiết kế để tối đa hóa thời gian người dùng.

+ Một hệ thống chấm điểm tín dụng không "*phân biệt đối xử*", nhưng lại tái tạo thiên lệch xã hội có sẵn trong dữ liệu.
+ Một mô hình ngôn ngữ tạo ra thông tin sai lệch không "*nói dối*", nhưng vẫn có thể gây hậu quả nghiêm trọng.

Những ví dụ này cho thấy: AI không phải là chủ thể đạo đức, nhưng lại là công cụ khuếch đại đạo đức hoặc phi đạo đức của con người.

Từ đó, vấn đề cần đặt ra không phải là "*làm sao để AI có đạo đức*", mà là: làm sao để con người không đánh mất đạo đức khi sử dụng AI.

Giữ giới trong kỷ nguyên AI

Thế giới hiện nay đang bước vào giai đoạn công nghệ phát triển nhanh hơn khả năng thích ứng về đạo đức. AI không chỉ thay đổi cách chúng ta làm việc, mà còn âm thầm định hình lại cách chúng ta suy nghĩ, tương tác và nhận thức về thực tại.

Trong bối cảnh đó, các hệ thống pháp lý và nguyên tắc đạo đức hiện đại, dù cần thiết, vẫn thường mang tính phản ứng, tức là chỉ xuất hiện sau khi vấn đề đã xảy ra. Ngũ giới của Phật giáo mang tính phòng hộ từ gốc, hướng đến việc chuyển hóa con người trước khi hành vi phát sinh.

Điều đáng chú ý là: Ngũ giới không đưa ra những quy định phức tạp, mà chỉ là năm nguyên tắc rất căn bản. Nhưng chính sự giản dị đó lại tạo nên sức sống lâu dài, bởi Ngũ giới chạm đến những vấn đề cốt lõi của con người: sự sống, sở hữu, dục vọng, lời nói và ý thức.

Trong một thế giới nơi AI có thể: tạo ra thông tin không có thật, tái cấu trúc các mối quan hệ và thậm chí tham gia vào quyết định sinh tử, thì việc quay trở lại với những nguyên tắc đạo đức căn bản không phải là lạc hậu, mà là hình thức tỉnh thức sâu sắc.

Ở góc nhìn khác, AI có thể được xem như *"tấm gương phóng đại"* phản chiếu tâm thức con người. Nếu con người còn tham lam, sân hận và si mê, thì AI sẽ khuếch đại những điều đó với tốc độ và quy mô chưa từng có.

Do đó, câu hỏi quan trọng nhất không phải là: AI sẽ trở thành gì trong tương lai?

Mà là: Con người sẽ trở thành gì khi sống cùng AI?

Và ở điểm này, Ngũ giới không chỉ là những nguyên tắc đạo đức, mà có thể trở thành nền tảng định hướng cho nhân loại trong kỷ nguyên công nghệ.

Câu hỏi gợi mở dành cho tất cả chúng ta:

+ Khi chúng ta trao quyền quyết định cho AI, liệu đó có phải là sự tiến bộ, hay là sự thoái lui về trách nhiệm đạo đức?

+ Nếu AI ngày càng "*giống người*", liệu chúng ta có đang đánh mất những phẩm chất làm nên chính mình?

+ Trong thế giới nơi sự thật có thể bị tái tạo vô hạn, đâu là nền tảng để thiết lập niềm tin?

+ Cuối cùng: giữ giới trong thời đại AI - là giới hạn, hay là tự do?

Tác giả: **Thường Nguyên**

Chú thích mở rộng: Những cảnh báo từ thực tiễn AI

[1] Cambridge Analytica scandal (2018)

Vụ bê bối Cambridge Analytica scandal là một trong những sự kiện chấn động nhất trong lịch sử công nghệ hiện đại, liên quan đến việc khai thác trái phép dữ liệu cá nhân của hàng chục triệu người dùng Facebook nhằm phục vụ mục đích chính trị.

Sự kiện chính:

+ Thời điểm: 2018
+ Quy mô dữ liệu: Khoảng 87 triệu tài khoản Facebook bị thu thập trái phép
+ Tổ chức liên quan: Cambridge Analytica (công ty con của SCL Group)

+ Người tố giác: Christopher Wylie
+ Hệ quả: Facebook bị Ủy ban Thương mại Liên bang Mỹ (FTC) phạt 5 tỷ USD; giá trị thị trường sụt giảm mạnh (mất hơn 100 tỷ USD).

Diễn biến vụ việc: Năm 2014, nhà nghiên cứu Aleksandr Kogan phát triển ứng dụng "*thisisyourdigitallife*" dưới danh nghĩa nghiên cứu học thuật, thu thập dữ liệu người dùng và cả bạn bè của họ. Dữ liệu này sau đó được chuyển giao cho Cambridge Analytica mà không có sự đồng thuận đầy đủ.

Công ty này sử dụng các mô hình tâm lý học hành vi (psychographic profiling - tạm dịch: phân tích hồ sơ tâm lý) để xây dựng các chiến dịch quảng cáo chính trị vi mô (micro-targeting), nhằm tác động đến hành vi cử tri trong cuộc bầu cử Tổng thống Mỹ năm 2016 và chiến dịch Brexit tại Anh.

Vụ việc được phanh phui năm 2018 qua điều tra của The Observer và The New York Times, làm dấy lên làn sóng tranh luận toàn cầu về quyền riêng tư dữ liệu.

Phản ứng và điều tra: Sau khi bê bối bùng nổ, các cơ quan quản lý tại Mỹ, Anh và châu Âu đồng loạt mở điều tra. CEO Mark Zuckerberg đã phải điều trần trước Quốc hội Mỹ và Nghị viện châu Âu.

Cambridge Analytica tuyên bố phá sản vào tháng 05/2018. Facebook sau đó cam kết cải tổ chính sách quyền riêng tư, siết chặt quyền truy cập dữ liệu của các ứng dụng bên thứ ba.

Ảnh hưởng lâu dài: Sự kiện này trở thành bước ngoặt trong nhận thức toàn cầu về quyền dữ liệu cá nhân (data privacy). Nó góp phần thúc đẩy việc triển khai Quy định bảo vệ dữ liệu chung của châu Âu (GDPR - General Data Protection Regulation, tạm dịch: Quy định chung về bảo vệ dữ liệu).

Đồng thời, vụ việc cũng làm nổi bật khái niệm "*chủ nghĩa tư bản giám sát*" (surveillance capitalism - tạm dịch: chủ nghĩa tư bản dựa trên giám sát dữ liệu), cảnh báo về việc dữ liệu cá nhân có thể bị khai thác như một công cụ thao túng quyền lực nếu thiếu kiểm soát minh bạch.

Dù Cambridge Analytica đã giải thể, vụ bê bối vẫn được xem là biểu tượng của những rủi ro đạo đức trong kỷ nguyên số, nơi dữ liệu cá nhân không chỉ là tài sản, mà còn là công cụ định hình nhận thức và hành vi xã hội.

*** Nguồn tham khảo thông tin:**

https://cambridgeanalytica.org/guides/how-the-cambridge-analytica-scandal-changed-the-internet-forever-3513/

https://www.dw.com/en/facebooks-cambridge-analytica-data-scandal-what-you-need-to-know/a-43071390
https://www.theguardian.com/uk-news/2018/may/03/cambridge-analytica-closing-what-happened-trump-brexit
https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook

[2] Một trong những vụ việc tiêu biểu cho sự lạm dụng công nghệ deepfake (giả mạo hình ảnh/video bằng AI) là trường hợp Taylor Swift vào đầu năm 2024.

Các hình ảnh khiêu dâm giả mạo bằng AI của cô đã lan truyền nhanh chóng trên mạng xã hội X (Twitter), với một số nội dung đạt tới hàng chục triệu lượt xem trước khi bị gỡ bỏ.

Dù là hình ảnh "*không có thật*", nhưng hậu quả lại hoàn toàn có thật: Xâm phạm nghiêm trọng danh dự và nhân phẩm. Gây tổn hại tâm lý kéo dài. Khó kiểm soát và gần như không thể xóa hoàn toàn khỏi Internet.

Các chuyên gia nhận định rằng phần lớn nội dung deepfake hiện nay là khiêu dâm không đồng thuận (non-consensual pornography - tạm dịch: nội dung khiêu dâm không có sự đồng thuận), và hiện tượng này đang trở thành một dạng bạo lực kỹ thuật số nhắm chủ yếu vào phụ nữ.

Vụ việc đã thúc đẩy các nhà lập pháp Mỹ đề xuất dự luật nhằm hình sự hóa việc phát tán nội dung deepfake khiêu dâm, cho thấy mức độ nghiêm trọng của vấn đề trong bối cảnh AI phát triển nhanh chóng.

***Nguồn tham khảo thông tin:**
https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box

[3] Luật sư sử dụng AI tạo thông tin sai trong hồ sơ pháp lý

Một trường hợp điển hình cho hiện tượng ảo giác AI (AI hallucination - tạm dịch: AI tạo thông tin sai nhưng trình bày như thật) xảy ra trong lĩnh vực pháp lý.

Năm 2025, hai luật sư tại Mỹ đã sử dụng AI để hỗ trợ soạn thảo hồ sơ, nhưng các án lệ được trích dẫn trong văn bản lại hoàn toàn không tồn tại. Tòa án sau đó xác định đây là thông tin bịa đặt và đã xử phạt luật sư số tiền 3.000 USD.

Vụ việc được VnExpress tường thuật lại, trở thành một ví dụ điển hình cho rủi ro khi con người phụ thuộc vào AI mà thiếu kiểm chứng.

Ý nghĩa đặc biệt của trường hợp này: AI không "*nói dối*" theo nghĩa đạo đức, nhưng tạo ra thông tin sai như thể là thật. Người sử dụng AI vẫn là chủ thể chịu trách nhiệm. Đặt ra yêu cầu cấp thiết về kiểm chứng thông tin (verification – tạm dịch: xác minh) trong thời đại AI.

***Nguồn tham khảo thông tin:** https://vnexpress.net/luat-su-bi-phat-3-000-usd-vi-dung-ai-viet-ho-so-day-loi-4911956.html

Tài liệu tham khảo:

1] Nguồn học thuật quốc tế về đạo đức AI
+ Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.
+ Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30, 99-120.
+ Floridi, L., et al. (2018). AI4People An Ethical Framework for a Good AI Society. Minds and Machines, 28, 689-707.
+ Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1, 501-507.

2] Nghiên cứu về Phật học và đạo đức
+ Harvey, P. (2000). An Introduction to Buddhist Ethics: Foundations, Values and Issues. Cambridge University Press.
+ Keown, D. (2005). Buddhist Ethics: A Very Short Introduction. Oxford University Press.
+ Bodhi, Bhikkhu (Ed.). (2012). The Numerical Discourses of the Buddha: A Translation of the Aṅguttara Nikāya. Wisdom Publications.
+ Gethin, R. (1998). The Foundations of Buddhism. Oxford University Press.

3] Nghiên cứu giao thoa Phật giáo và AI
+ Nguyen, T. C. (2024). The Five Precepts and Digital Ethics in the Age of Artificial Intelligence.
+ MDPI. (2025). Buddhist Ethics and Artificial Intelligence: Toward Compassionate AI Systems. Religions Journal.
+ Journal of Buddhist Innovation Review. (2025). Ethical Reflections on AI through Buddhist Thought.

4] Nguồn báo chí và thực tiễn
+ BBC (2023-2025). Các bài viết về chiến tranh hiện đại, AI và vũ khí tự hành.
+ Reuters (2023–2025). Báo cáo về xung đột tại Ukraine, Gaza và ứng dụng công nghệ quân sự.
+ The New York Times (2018). How Cambridge Analytica Exploited Facebook

Data.
+ The Guardian (2023-2025). Các bài viết về deepfake, AI và khủng hoảng thông tin.
+ VnExpress (2023-2025). Các bài viết về nghiện mạng xã hội, dữ liệu cá nhân và công nghệ.

5] Sự kiện và khái niệm tiêu biểu
+ Cambridge Analytica scandal
+ European Commission. (2019). Ethics Guidelines for Trustworthy AI.
+ UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence.

VIII. Phụ lục thuật ngữ và diễn giải:

Thuật ngữ tiếng Anh	Tạm dịch đề xuất	Ghi chú học thuật
AI ethics	đạo đức AI	dùng phổ biến
transparency	tính minh bạch	
fairness	tính công bằng	
accountability	trách nhiệm giải trình	rất quan trọng
privacy	quyền riêng tư	
non-maleficence	nguyên tắc không gây hại	gốc từ y đức
misinformation	thông tin sai lệch	không nhất thiết có chủ ý
disinformation	thông tin sai lệch có chủ đích	nên phân biệt
AI hallucination	ảo giác AI	AI "bịa" nhưng tưởng đúng
bias (algorithmic bias)	thiên lệch (thuật toán)	
autonomous weapons	vũ khí tự hành	liên quan giới 1
surveillance	giám sát	

Ngũ giới	Giới hạn đạo đức của AI
Không sát sinh	Vấn đề vũ khí tự hành (autonomous weapons) và hệ thống gây hại đến sự sống
Không trộm cắp	Vi phạm quyền riêng tư (privacy), khai thác dữ liệu trái phép
Không tà dâm	Lạm dụng AI trong nội dung nhạy cảm, deepfake và robot tình dục
Không nói dối	Thông tin sai lệch (misinformation), ảo giác AI (AI hallucination)
Không sử dụng chất gây nghiện	Nghiện công nghệ, thiết kế gây lệ thuộc (addictive design)