WHEN THE ALGORITHM GETS IT WRONG

Sai lầm của thuật toán

Tác hại của AI và các vấn đề đạo đức qua hồ sơ AIAAIC

Những gì đã xảy ra trong tuyển dụng, phúc lợi, tòa án, đường phố, lớp học và nơi làm việc

Kim Kyung-jin, Luật sư

Mục lục

CHƯƠNG 1

Thiên kiến và phân biệt đối xử: Định kiến của con người được viết vào mã

1.1

Thiên lệch thuật toán trong tuyển dụng, tài chính và dịch vụ

Những email từ chối luôn đến vào lúc rạng sáng.

Derek Mobley là một người đàn ông da đen ngoài bốn mươi tuổi, đang mắc chứng rối loạn lo âu và trầm cảm. Ông làm việc trong lĩnh vực công nghệ thông tin, có chứng chỉ và cũng có kinh nghiệm. Ông đã nộp đơn ứng tuyển vào khoảng một trăm công ty sử dụng hệ thống tuyển dụng của một công ty tên là Workday.

Ông đã trượt khoảng một trăm lần.

Nhìn vào thời gian email từ chối được gửi đến, có thể biết sự thật rằng đó không phải do con người gửi. Một giờ sáng, ba giờ sáng. Người phụ trách tuyển dụng đang ngủ vào giờ đó. Máy móc thì không ngủ. Chúng không uống cà phê, cũng không biết xin lỗi. Chúng mở hồ sơ ra và đóng lại chỉ trong vòng ba giây.

Mobley đã đệ đơn kiện. Lập luận của ông là trí tuệ nhân tạo của Workday hoạt động giống như một cái rây lọc độ tuổi, chủng tộc và khuyết tật. Tòa án đã chấp thuận một vụ kiện tập thể quy mô toàn quốc sau khi nhận định rằng Workday đã huấn luyện trí tuệ nhân tạo bằng dữ liệu của các nhân viên hiện tại, qua đó sử dụng các đặc điểm cần được bảo vệ như một chỉ số đại diện.

Tháng 3 năm 2026, tòa án đã bác bỏ lập luận của phía Workday rằng Đạo luật chống phân biệt đối xử vì tuổi tác không áp dụng cho các ứng viên. Vào ngày 16 tháng 6, đã có phán quyết rằng Workday có thể phải chịu trách nhiệm ngay cả khi nhà tuyển dụng ở ngoài California sử dụng hệ thống. Số lượng hồ sơ bị từ chối mà vụ kiện này đang xử lý là một tỷ một trăm triệu hồ sơ.

Đó là một tỷ một trăm triệu email rạng sáng.

Lý do khiến câu chuyện này đáng sợ nằm ở chỗ không hề có kẻ ác. Không có ai tụ tập trong phòng họp và quyết định đánh trượt những người ngoài bốn mươi tuổi. Họ chỉ đơn giản là nạp cho máy móc dữ liệu của những người đã được tuyển dụng trong mười năm qua. Máy móc đã học các khuôn mẫu đó một cách chăm chỉ. Và nó cũng lặp lại một cách chăm chỉ.

Từ học máy nghe có vẻ khó hiểu nhưng nguyên lý thì giống với giờ học toán ở trường tiểu học. Họ không cho biết quy tắc mà chỉ đưa ra rất nhiều đáp án, rồi bắt tìm ra quy tắc từ đó. Nếu những đáp án trong quá khứ mang tính thiên lệch, máy móc sẽ học sự thiên lệch đó như một quy tắc. Từ góc độ của máy móc, nó đã làm tốt công việc của mình. Bởi vì nó đã làm đúng như được sai bảo.

Mười năm trước, Amazon đã học được sự thật đó đầu tiên.

Năm 2014, một dự án bí mật đã được bắt đầu tại trung tâm phát triển của Amazon ở Edinburgh, Scotland. Đó là một hệ thống chấm điểm đánh giá bằng sao khi người ta nộp sơ yếu lý lịch. Từ một sao đến năm sao. Có thể coi đó là việc chấm điểm bằng sao cho con người giống như chấm điểm cho hàng hóa.

Các nhà phát triển đã đưa hồ sơ ứng tuyển vào các vị trí kỹ thuật trong mười năm qua vào làm dữ liệu huấn luyện. Trong suốt mười năm đó, các nhân sự kỹ thuật mà Amazon đã tuyển dụng hầu hết là nam giới. Máy móc đã đưa ra kết luận. Nam giới là những ứng viên tốt hơn.

Khi tìm thấy từ phụ nữ trong sơ yếu lý lịch, máy móc sẽ trừ điểm. Những sơ yếu lý lịch ghi là chủ tịch câu lạc bộ cờ vua nữ đã bị trừ điểm. Điểm số của những ứng viên tốt nghiệp từ các trường đại học nữ cũng bị đánh tụt xuống toàn bộ. Ngược lại, nếu hồ sơ có chứa những động từ mà các kỹ sư nam thường dùng, như "đã thực hiện" hoặc "đã nắm bắt", điểm số sẽ tăng lên.

Amazon đã cố gắng sửa chữa sự thiên lệch này. Họ đã thất bại. Năm 2017, họ đã hủy bỏ dự án. Thật may mắn, Amazon tuyên bố rằng hệ thống đó đã không được sử dụng trong việc tuyển dụng thực tế.

Vấn đề nằm ở chỗ các công ty khác đã liên tục tạo ra những thứ tương tự.

Kyle Behm là một sinh viên của Đại học Vanderbilt. Anh đã tạm nghỉ học để điều trị chứng rối loạn lưỡng cực, và đã ứng tuyển vào các cửa hàng ở địa phương để kiếm tiền sinh hoạt. Siêu thị Kroger, Home Depot, Walgreens, PetSmart. Đó là những công việc thu ngân bán thời gian.

Anh ấy đã bị loại khỏi tất cả. Anh thậm chí còn không đi được đến vòng phỏng vấn.

Nguyên nhân là do bài kiểm tra tính cách trực tuyến tên là Unicru do một công ty tên là Kronos tạo ra đã bật đèn đỏ đối với Kyle. Bài kiểm tra đó đã phát hiện ra tiền sử sức khỏe tâm thần của Kyle và phán đoán rằng một người như vậy sẽ có khả năng cao phớt lờ khách hàng khi khách hàng nổi giận.

Cha của Kyle, Roland Behm, là một luật sư. Năm 2014, ông đã đệ đơn kiến nghị lên Ủy ban Cơ hội Việc làm Bình đẳng, và nộp đơn kiện các doanh nghiệp về tội phân biệt đối xử với người khuyết tật tâm thần.

Kyle đã tự kết liễu đời mình vào năm 2019. Đó là trước khi vụ kiện kết thúc.

Sự thật rằng đã có một chàng trai trẻ không thể đứng ở quầy thu ngân vì không vượt qua được bài kiểm tra tính cách. Sự thật rằng thứ đưa ra phán quyết đó không phải là con người mà là một trang web. Và sự thật rằng không có kênh khiếu nại nào trên trang web đó. Ghép ba câu này lại với nhau và đọc, chúng ta sẽ thấy lạnh sống lưng.

Hệ thống của iTutorGroup còn trắng trợn hơn. Chương trình tuyển dụng tự động của công ty giáo dục trực tuyến này đã tự động loại bỏ phụ nữ trên năm mươi lăm tuổi và nam giới trên sáu mươi tuổi. Hơn hai trăm giáo viên đã bị sa thải vì lý do tuổi tác mặc dù họ có đủ trình độ.

Cách thức mà việc này bị phanh phui lại rất nực cười. Một ứng viên đã nộp đơn với ngày tháng năm sinh thật và bị loại. Nhưng khi giảm độ tuổi xuống và nộp đơn lại, người đó ngay lập tức nhận được liên lạc mời đi phỏng vấn. Nếu là con

người thì chắc hẳn đã đỏ mặt, nhưng hệ thống thì vẫn bình thản. Ủy ban Cơ hội Việc làm Bình đẳng đã phạt số tiền bồi thường hòa giải là ba trăm sáu mươi lăm nghìn đô la.

Tôi đã nghĩ rằng mọi chuyện sẽ tốt hơn khi trí tuệ nhân tạo tạo sinh ra đời.

Năm 2024, nhóm nghiên cứu của Bloomberg đã cung cấp cho GPT-3.5 rất nhiều sơ yếu lý lịch giả với trình độ chuyên môn y hệt nhau và chỉ thay đổi tên. Đó là những cái tên có thể giúp đoán được chủng tộc và giới tính. Kết quả là như sau. Trong việc đánh giá các vị trí kỹ thuật, những cái tên được cho là của phụ nữ da đen được chọn ít hơn 36 phần trăm. Tên của phụ nữ bị đẩy sang các công việc quản lý nhân sự.

Năm 2021, đài truyền hình Bayern của Đức đã thử nghiệm hệ thống đánh giá năng lực bằng trí tuệ nhân tạo của một công ty khởi nghiệp tên là Retorio. Cùng một người nói cùng một câu nhưng khi đeo kính, điểm số lại thay đổi. Khi đội khăn trùm đầu hijab, điểm số lại thay đổi lần nữa. Nếu nhìn thấy tủ sách ở phía sau, hệ thống đưa ra phán quyết rằng người đó trông có vẻ chăm chỉ.

Thật nực cười khi tính cách lại trở nên tốt hơn chỉ nhờ đặt một cái tủ sách. Tiếng cười sẽ dừng lại ở chi tiết rằng bằng phán quyết đó, có người có được việc làm và có người lại không có được.

Điều tương tự cũng xảy ra trên bục giảng. Năm 2012, Sarah Wysocki, một giáo viên lớp 5 tại một trường công lập ở Washington, D.C., đã bị sa thải bởi một hệ thống đánh giá giáo viên tên là IMPACT. Đánh giá của các đồng nghiệp và phụ huynh học sinh đều rất xuất sắc. Tuy nhiên, đã từng có tình trạng gian lận thi cử ở ngôi trường trước đây mà các học sinh theo học, và những điểm số bị thổi phồng do việc đó đã kéo tụt điểm giá trị gia tăng của Wysocki.

Hai trăm lẻ năm giáo viên đã mất việc chỉ vì một công thức toán học chưa được kiểm chứng. Công thức đó là gì thì chưa từng được công bố.

Các thuật toán xử lý tiền bạc đã làm việc lâu hơn và lặng lẽ hơn.

Năm 2018, các nhà nghiên cứu từ Đại học California Berkeley đã công bố một báo cáo. Họ phân tích những khoản vay lãi suất cố định 30 năm được bảo đảm bởi Fannie Mae và Freddie Mac từ năm 2008 đến 2015. Lúc đó, người ta thường nói rằng nếu chuyển việc phê duyệt vay từ con người sang thuật toán, sự phân biệt sẽ biến mất.

Nhưng con số lại nói khác. Những người vay tiền là người da đen và Latinh đã phải trả thêm 5,6 đến 8,6 điểm cơ bản cho khoản vay mua nhà, và 3 điểm cơ bản cho tái cấp vốn. Điểm cơ bản là một đơn vị tài chính có nghĩa là 0,01 phần trăm. Đó là những con số tưởng chừng như nhỏ.

Khi cộng những con số nhỏ này trên toàn bộ nước Mỹ, con số đó trở thành 250 triệu đến 500 triệu đô la mỗi năm.

Lý do lại càng tủi hơn. Các tổ chức cho vay đã dùng trí tuệ nhân tạo để phát hiện ra rằng những người vay từ các dân tộc thiểu số không có đủ thời gian và thông tin để so sánh nhiều sản phẩm. Rồi họ định giá cao hơn cho những người đó. Điều này gọi là định giá chiến lược. Lợi nhuận bổ sung mà họ kiếm được bằng cách đó là từ 11 phần trăm đến 17 phần trăm.

Tờ báo điều tra The Markup và hãng tin Yonhap News cùng phân tích dữ liệu vay mua nhà liên bang. Một thuật toán phê duyệt bí mật sử dụng điểm FICO cổ điển đã từ chối những đơn xin cấp vốn từ các ứng viên da màu nhiều hơn 40 đến 80 phần trăm so với những ứng viên da trắng. Ở một số thành phố lớn, con số đó vượt quá 250 phần trăm.

Laura Blattner và Scott Nelson từ Stanford Graduate School of Business đã đi sâu hơn nữa. Họ chỉ ra rằng ngay cả khi thuật toán không có ý định xấu, vấn đề vẫn có thể phát sinh. Những người có thu nhập thấp và những người từ các dân tộc thiểu số có lịch sử tín dụng mỏng manh. Khi lịch sử mỏng, độ chính xác dự báo sẽ giảm từ 5 đến 10 phần trăm. Một hoặc hai khoản nợ quá hạn từ lâu có thể làm hỏng toàn bộ điểm số.

Nếu bạn nghèo, dữ liệu sẽ ít; nếu dữ liệu ít, máy không thể dự đoán chính xác; nếu không dự đoán chính xác, bạn sẽ được phân loại là người có rủi ro cao. Và cái phân loại đó được gắn nhãn là "rủi ro tín dụng khách quan".

Tại ngân hàng Wells Fargo, một sai sót tính toán đã đẩy mọi người ra đường. Từ 2010 đến 2015, phần mềm xem xét tái điều chỉnh thế chấp tự động của ngân hàng này có lỗi tính toán sai chi phí luật sư. Điều này suốt năm năm mà không ai phát hiện ra.

Tám trăm bảy mươi người đủ điều kiện để thay đổi khoản vay của họ đã bị từ chối. Trong số đó, ít nhất năm trăm bốn mươi lăm người đã mất nhà do bị tịch biên. Ngân hàng sửa lỗi vào cuối năm 2015, nhưng không công bố điều này cho công chúng trong vài năm. Khoản bồi thường được xác định sau đó là tám triệu đô la.

So với giá trị một ngôi nhà, đó là một khoản tiền không tương xứng.

Bảo hiểm cũng vậy. Các thuật toán đánh giá rủi ro được sử dụng bởi GEICO, Safeco và Nationwide đã áp dụng phí bảo hiểm cao hơn tối đa 30 phần trăm cho những cư dân sống ở những khu vực có nhiều người da màu, mặc dù rủi ro tai nạn là như nhau. Allstate thậm chí còn làm nhiều hơn. Nó tăng giá nếu khách hàng có vẻ không muốn rời đi. Những khách hàng trung thành giao dịch lâu đã phải trả giá cao hơn. Cơ quan quản lý đã phát hiện ra điều này.

Ngay cả khi vào nơi làm việc, thuật toán cũng tiếp tục theo đuổi bạn.

Năm 2021, Tòa án Lao động Bologna ở Ý đã phán quyết rằng thuật toán Frank của nền tảng giao hàng Deliveroo là phân biệt. Frank đưa ra điểm số cho độ tin cậy và mức độ tham gia của những công nhân giao hàng. Nếu họ không xuất hiện trong ca làm việc đã định sẵn, điểm số của họ sẽ bị trừ. Nó không hỏi lý do.

Nếu ốm và không thể đến, điểm số cũng bị trừ. Nếu tham gia đình công được bảo vệ bởi pháp luật, điểm số cũng bị trừ. Khi điểm số bị trừ, họ không nhận được những đơn giao hàng tốt lần tới. Quyền đình công được ghi lại trong pháp

luật, nhưng cấu trúc của mã máy tính đã tạo ra một tình huống trong đó tham gia đình công sẽ dẫn đến đói kém.

Hệ thống xác minh danh tính tự động của một công ty tên là Checkr đã lọc ra hàng loạt những người lái xe của Uber, Lyft và Postmates từ năm 2015. Những hồ sơ về tội phạm của những người khác cùng tên, những hồ sơ đã hết hạn thời hiệu, đều được gắn vào như vậy. Những tài xế mà tài khoản bị khóa ngay một ngày không biết phải phản đối ở đâu.

Giáo sư Veena Dubal từ Trường Luật Hastings, Đại học California, đã đặt tên cho hiện tượng này. Đó là phân biệt đối xử trong lương thưởng do thuật toán quyết định. Nó có nghĩa là Uber, Lyft và Amazon thu thập dữ liệu về công nhân để trả những khoản tiền khác nhau cho cùng một công việc. Vì bạn không biết người bên cạnh nhận bao nhiêu, bạn không thể so sánh.

Hành khách cũng không thoải mái. Theo nghiên cứu, thuật toán định giá và phân bổ xe của Uber và Lyft đã áp dụng những khoản phí cao hơn và thời gian chờ lâu hơn cho những khu vực có nhiều cư dân da màu.

Hệ thống quảng cáo Facebook đã lọc các thông báo tuyển dụng dựa trên tuổi tác. Verizon, Amazon và Goldman Sachs đã sử dụng tính năng nhắm mục tiêu theo tuổi để ẩn các thông báo với những người tìm kiếm việc làm lớn tuổi. Một nghiên cứu năm 2021 của Đại học Northeastern và một phán quyết năm 2025 của Nhân viên Bảo vệ Quyền Con người Pháp đã đi xa hơn nữa. Ngay cả khi nhà quảng cáo không chỉ định giới tính, thuật toán phân phát nội bộ của Facebook vẫn tự động cho những người đàn ông và phụ nữ xem những thông báo tuyển dụng khác nhau.

Giá của đồ vật cũng được xác định dựa trên bộ mặt của người mua.

Định giá thông minh của Airbnb là một tính năng tự động điều chỉnh giá tiền phòng. Khi các nhà nghiên cứu do Giáo sư Param Vir Singh từ Đại học Carnegie Mellon dẫn đầu xem xét, họ phát hiện ra rằng những chủ nhà da trắng sử dụng tính năng này nhiều hơn, và kết quả là khoảng cách thu nhập giữa những chủ

nhà da trắng và da đen lại còn tăng thêm.

Học phí lớp học trực tuyến của Princeton Review thay đổi tùy theo mã bưu điện. Những khu vực có nhiều người Mỹ gốc Á được áp dụng giá cao hơn. Báo chí gọi điều này là "Thuế Tiger Mom".

Trang web du lịch Orbitz đã hiển thị những khách sạn đắt tiền ở vị trí cao hơn cho những người dùng máy tính Mac. Điều này là vì số liệu thống kê cho thấy người dùng Mac có thu nhập trung bình cao hơn. Nó đã phán quyết rằng nếu nắp máy tính xách tay có hình quả táo, ví tiền sẽ dày hơn.

Amazon có một thuật toán bí mật gọi là Project Nessie. Nó là một thử nghiệm trong đó công ty nhẹ nhàng nâng giá sản phẩm của riêng mình để xem liệu đối thủ cạnh tranh có theo sau không. Thuật toán Buy Box đã cho người tiêu dùng thấy những sản phẩm mang lại nhiều hoa hồng cho Amazon trước tiên, thay vì cho họ thấy sản phẩm rẻ nhất.

Tất cả các trường hợp này có một điểm chung. Khi vấn đề phát sinh, điều mà các công ty nói đều gần như giống nhau.

Đó là kết quả của thuật toán tự học. Đó là một lỗi kỹ thuật. Cấu trúc bên trong là bí mật thương mại nên không thể công bố được.

Ba câu đã đủ. Người chọn dữ liệu, người thiết lập hàm mục tiêu, và người hưởng lợi từ kết quả đều ngồi trong phòng họp, nhưng trách nhiệm lại rơi vào mã máy tính. Mã máy tính không biện hộ. Nó cũng không xin lỗi. Nó cũng không xuất hiện tại tòa án.

Không có nhân viên nào thuận tiện hơn thế này trên thế giới.

1.2

Lỗi thuật toán trong việc nhận phúc lợi và mạng lưới an toàn xã hội

Bức thư trông có vẻ bình thường. Nó được đặt trong phong bì của chính phủ Úc, với phông chữ của chính phủ và được đóng dấu của chính phủ. Bên trong có viết: Quý vị đã nhận trợ cấp phúc lợi vượt mức. Vui lòng hoàn trả số tiền dưới đây.

Những người nhận được thư đã vô cùng ngạc nhiên khi nhìn thấy số tiền. Có những bức thư ghi số tiền lên tới hơn mười triệu won. Và họ còn ngạc nhiên lần thứ ba. Không có bất kỳ lời giải thích nào về lý do tại sao họ phải trả lại tiền.

Có hơn bốn trăm ngàn người đã nhận được bức thư này.

Đây là việc làm của hệ thống can thiệp tuân thủ trực tuyến mà Bộ Dịch vụ Xã hội Úc và Centrelink đã vận hành từ năm 2016 đến năm 2019, một chương trình mà mọi người gọi là Robodebt. Robo là rô-bốt, còn debt là nợ. Điều này có nghĩa là món nợ do rô-bốt tạo ra.

Lỗi mà hệ thống này gây ra chỉ ở mức độ toán cấp một.

Chính phủ đã đối chiếu dữ liệu thu nhập hàng năm của cơ quan thuế với hồ sơ trợ cấp phúc lợi. Vấn đề nằm ở chỗ đơn vị thời gian của hai dữ liệu này khác nhau. Cơ quan thuế có dữ liệu gộp của cả một năm, trong khi trợ cấp phúc lợi được chi trả hai tuần một lần.

Hệ thống đã chọn một cách làm dễ dàng. Nó chia thu nhập một năm cho năm mươi hai và giả định rằng mỗi tuần người đó kiếm được số tiền bằng nhau.

Nếu là nhân viên công sở làm việc toàn thời gian thì phép tính này là chính xác. Tuy nhiên, những người nhận trợ cấp phúc lợi phần lớn không phải là nhân viên toàn thời gian. Họ là những người làm việc ba tháng vào mùa hái dâu và nghỉ

ngơi trong thời gian còn lại, những người tháng này làm hai công việc nhưng tháng sau không có việc nào, hay những người phải nộp đơn xin trợ cấp lại mỗi khi hết hạn hợp đồng.

Nếu chia thu nhập một năm của những người này cho năm mươi hai thì sẽ ra sao? Kết quả sẽ tính là họ có kiếm được tiền ngay cả trong những tuần không làm việc. Do đó, hệ thống đưa ra kết luận rằng họ đã nhận trợ cấp một cách bất hợp pháp.

Thông báo đòi nợ đã được gửi đến những người không hề mắc nợ.

Để khiếu nại, họ phải mang đến bảng lương từ vài năm trước. Nếu không có, món nợ đó sẽ thuộc về họ. Đó là một cơ cấu mà công dân phải tự chứng minh sự trong sạch của mình chứ không phải chính phủ chứng minh phép tính. Các công ty thu hồi nợ đã vào cuộc. Tiền hoàn thuế bị tịch thu.

Một số người đã không thể chịu đựng được. Đã có những người tự kết liễu đời mình vì không tìm được nơi để giải tỏa nỗi oan ức.

Tại phiên điều trần của Ủy ban Điều tra Hoàng gia năm 2023, một sự thật vô cùng đáng sợ đã được tiết lộ. Đó là đã có nhiều cảnh báo rằng hệ thống này là bất hợp pháp và phương pháp tính toán có lỗ hổng. Các chuyên gia pháp lý đã cảnh báo, và cả các nhân viên nội bộ cũng đã cảnh báo.

Các quan chức cấp cao và bộ trưởng đều biết điều đó. Thế nhưng họ vẫn tiếp tục vận hành nó.

Đó là bởi vì họ cần thành tích tiết kiệm ngân sách. Nếu con người xác nhận từng trường hợp một thì sẽ mất rất nhiều thời gian và tiền bạc. Việc loại bỏ quy trình xác nhận đó chính là chức năng cốt lõi của hệ thống này. Kết cục là một thỏa thuận trị giá một tỷ tám trăm triệu đô la Úc và lời xin lỗi chính thức từ chính phủ.

Lời xin lỗi đã không thể đến được với những người đã khuất.

Ở Ấn Độ, ngón tay lại là vấn đề.

Chính phủ Ấn Độ quản lý phúc lợi bằng một hệ thống xác thực sinh trắc học có tên là Aadhaar. Người dân phải quét vân tay thì mới nhận được khẩu phần lương thực. Tại một ngôi làng hẻo lánh ở bang Jharkhand, phương pháp này đã giết chết con người.

Máy quét vân tay liên tục bị lỗi. Đầu ngón tay của những người làm việc trên cánh đồng và lao động chân tay cả đời đã bị bào mòn. Dấu vân tay trở nên mờ đi. Hơn nữa, ở những ngôi làng miền núi, tín hiệu internet rất kém. Việc chờ đợi tín hiệu đã mất đi một ngày.

Người dân thuộc các tầng lớp bị yếu thế, bao gồm cả người Parhaiya, đã bị từ chối phát khẩu phần lương thực được pháp luật bảo đảm. Ở nhiều ngôi làng, đã có những người chết vì đói hoặc chết vì suy dinh dưỡng.

Không phải là máy móc không nhận ra con người. Chính quy tắc quy định rằng không được phát thức ăn nếu máy móc không nhận dạng được mới là thứ đã giết chết con người.

Ở bang Haryana, một hệ thống có tên là Parivar Pehchan Patra đã xử lý hàng ngàn người còn sống như những người đã chết. Nếu bị ghi nhận là đã chết, lương hưu sẽ bị cắt và khẩu phần lương thực cũng bị cắt. Sau này khi điều tra, người ta phát hiện ra rằng 70 phần trăm các khoản lương hưu bị ngừng chi trả là do đánh giá sai của thuật toán.

Đã có những người cao tuổi phải đến các cơ quan chính quyền để chứng minh rằng mình còn sống.

Hệ thống Samagra Vedika của bang Telangana đã dự đoán sai về thu nhập và tình trạng việc làm, qua đó âm thầm xóa bỏ thẻ lương thực của mười lăm ngàn hộ gia đình. Phải đến khi người dân nổi dậy phản đối thì mọi chuyện mới được phục hồi.

Tại Jordan, một thuật toán đo lường nghèo đói có tên là Takaful do Ngân hàng Thế giới tài trợ đã được triển khai. Hệ thống này chấm điểm một vài đặc điểm của hộ gia đình để xác định đối tượng được hỗ trợ. Tuy nhiên, những gia đình thực sự đang đói khát lại bị loại bỏ do không đủ điểm. Những kẻ cố gắng đo lường cái nghèo bằng con số lại không thể nhận ra cái nghèo.

Ứng dụng Meu INSS của Brazil được tạo ra nhằm giảm thiểu hàng người chờ đợi nộp đơn xin lương hưu. Kết quả là hàng người đã giảm bớt. Đó là vì các đơn đăng ký đã bị từ chối tự động. Nếu có lỗi đánh máy trong biểu mẫu hoặc hiểu sai câu hỏi, họ sẽ bị loại ngay lập tức. Những người cao tuổi ở nông thôn không quen thuộc với điện thoại thông minh đã vướng phải tình trạng này. Để nộp đơn lại, họ phải thuê luật sư.

Các quốc gia giàu có cũng không hề khác biệt.

Universal Credit của Anh là một chế độ gộp nhiều loại trợ cấp phúc lợi lại thành một. Thuật toán tính toán được áp dụng ở đây đã quyết định mức trợ cấp dựa trên số tiền kiếm được trong một tháng theo lịch. Tổ chức Human Rights Watch đã điều tra cơ cấu này vào năm 2020.

Một số người lao động nhận lương bốn tuần một lần. Như vậy, sẽ có những tháng tiền lương được trả hai lần. Hệ thống đánh giá rằng người này đột nhiên trở nên giàu có. Trợ cấp của tháng đó đã bị cắt giảm mạnh hoặc hoàn toàn không được chi trả.

Tiền ăn đã biến mất tùy thuộc vào việc tờ lịch bắt đầu từ thứ mấy.

Thuật toán phát hiện nhận trợ cấp gian lận của Bộ Việc làm và Lương hưu Anh còn hoạt động âm thầm hơn. Một mô hình có tên là Advances và thuật toán giám sát tiền hỗ trợ nhà ở đã đánh giá điểm rủi ro dựa trên độ tuổi, quốc tịch, tình trạng khuyết tật và tình trạng hôn nhân của người nhận.

Danh sách những người nhận được điểm cao đã xuất hiện. Đó là những người phụ nữ lao động thu nhập thấp bán thời gian mang quốc tịch Bulgaria và Ba

Lan, cùng với những người khuyết tật.

Trong vòng ba năm, đã có hai trăm ngàn người bị phân loại vào nhóm nguy cơ cao, dẫn đến việc bị ngừng hỗ trợ hoặc bị điều tra. Hơn hai phần ba trong số đó là những người nhận trợ cấp hợp pháp, không hề có bất kỳ lỗi lầm nào.

Cơ quan phúc lợi UDK của Đan Mạch đã vận hành hơn sáu mươi thuật toán phát hiện nhận trợ cấp gian lận. Họ đã thu thập tình trạng cư trú, quốc tịch, mối quan hệ gia đình, hồ sơ y tế và cả lịch sử xuất cảnh. Quốc gia đã tập hợp lại tất cả những thông tin về việc người đó đã đi đâu, sống với ai và mắc bệnh gì để chấm điểm.

Các tổ chức nhân quyền quốc tế cho rằng đây là hệ thống tính điểm xã hội bị cấm bởi Luật AI của Liên minh Châu Âu. Tổ chức Amnesty International Quốc tế đã công bố một báo cáo có tiêu đề "Bất công Được Mã Hóa". Nội dung của nó cho rằng người khuyết tật, người có thu nhập thấp, người nhập cư, người tị nạn và các dân tộc thiểu số có nguy cơ cao bị đánh dấu là đối tượng điều tra.

Khi Amnesty yêu cầu được xem mã và dữ liệu, các cơ quan chính quyền Đan Mạch đã từ chối.

Cơ quan Bảo hiểm Xã hội Thụy Điển chạy mô hình dự đoán gian lận từ năm 2013. Khi Báo cáo Lighthouse xem xét kỹ lưỡng, phụ nữ, người nhập cư, người xuất thân nước ngoài, người có thu nhập thấp và người chưa tốt nghiệp đại học bị bắt gặp thường xuyên hơn.

Thuật toán đánh giá rủi ro của tổ chức bảo hiểm xã hội Pháp CNAF đã cho điểm gần một nửa dân số. Các gia đình một cha mẹ, người có thu nhập thấp và người khuyết tật tự động trở thành đối tượng nghi vấn. Họ không giải thích tại sao điểm lại được tính như vậy. Thay vào đó, họ đến để khám tìm nhà cửa.

Ở Hà Lan, cơ quan thuế quốc gia đã bước tới để bắt gian lận trợ cấp chăm sóc trẻ em từ năm 2013 đến 2019 bằng thuật toán tự học. Hàng ngàn phụ huynh bị coi là những kẻ gian lận. Quốc tịch kép được sử dụng làm tín hiệu rủi ro. Hệ

thống của Rotterdam phân chia người dân theo chủng tộc, tuổi, giới tính và liệu họ có con hay không.

Cũng có những thuật toán được tạo ra với danh nghĩa bảo vệ trẻ em.

Hạt Allegheny ở bang Pennsylvania, Hoa Kỳ đã sử dụng một công cụ sàng lọc gia đình để dự đoán rủi ro lạm dụng trẻ em. Khi có báo cáo đến, công cụ này cho điểm, và nếu điểm cao, người điều tra sẽ đến nhà.

Các nhà nghiên cứu tại Đại học Carnegie Mellon đã phân tích kết quả. Hai phần ba trẻ em da đen được báo cáo được khuyến cáo là đối tượng điều tra. Đối với trẻ em da trắng, nó là một nửa.

Lý do nằm ở những gì công cụ này đang xem xét. Công thức bao gồm lịch sử nhận được trợ cấp từ chính phủ. Nó cũng bao gồm các hồ sơ chẩn đoán sức khỏe tâm thần.

Nếu bạn nhận trợ cấp vì bạn nghèo, điểm rủi ro sẽ tăng lên. Nếu bạn nhận tư vấn vì bạn bị bệnh, điểm rủi ro sẽ tăng lên. Đó là một cấu trúc trong đó hồ sơ yêu cầu giúp đỡ trở thành căn cứ cho sự nghi ngờ.

Một chiếc cân mà càng nhận trợ giúp, xác suất bị tước con em càng cao. Bạn không thể tạo ra một chiếc cân như vậy và gọi nó là công bằng.

Hệ thống của Oregon được tạo bắt chước công cụ này đã bị loại bỏ vào năm 2022 sau khi thiên vị chủng tộc được phát hiện. Hệ thống Eckerd Rapid Safety Feedback mà Illinois đã sử dụng cũng bị ngừng vì lỗi dữ liệu và hiệu suất bị phóng đại.

Hệ thống hỗ trợ quyết định của Cơ quan Bảo vệ Trẻ em Đan Mạch được phát hiện có những thói quen lạ trong xác minh của Đại học Công nghệ Thông tin Copenhagen. Với tình huống giống nhau nhưng tuổi khác nhau, điểm rủi ro là khác nhau.

Chính phủ Úc đã cố gắng giới thiệu một công cụ đánh giá tự động gọi là Robo Planning để tiết kiệm 700 triệu đô la Úc từ ngân sách hỗ trợ người khuyết tật. Đó là một phương pháp nén cuộc sống của những người khuyết tật thành một vài loại tiêu chuẩn và sau đó tính toán sự hỗ trợ cần thiết. Các tổ chức mù lòa và các tổ chức người khuyết tật đã phản đối mạnh mẽ, và kế hoạch đã bị hoãn lại.

Trong bệnh viện, các thuật toán xác định ngày xuất viện.

nH Predict, được sử dụng bởi UnitedHealth Group và Humana, các công ty bảo hiểm sức khỏe lớn nhất ở Hoa Kỳ, dự đoán số ngày bệnh nhân cao tuổi và người khuyết tật nên nhận điều trị phục hồi chức năng. Ngay cả khi bác sĩ nói là cần thêm, khi ngày do hệ thống định rồi đến, bảo hiểm bị cắt.

Trong một vụ kiện, tỷ lệ lỗi từ chối yêu cầu của hệ thống này được công khai. Nó là 90 phần trăm.

Nếu từ chối mười lần, chín lần là sai. Tuy nhiên, lý do tiếp tục sử dụng nó rất đơn giản. Hầu hết bệnh nhân bị từ chối đã không kháng cáo. Sẽ khó khăn cho những người già và ốm yếu để chống lại các công ty bảo hiểm.

Vào tháng 11 năm 2021, Joan Barros, 86 tuổi, ngã và chấn thương chân. Nhân viên y tế nói rằng vật lý trị liệu là hoàn toàn cần thiết. Khi ngày xuất viện do thuật toán định rồi đến, bảo hiểm điều trị đã bị cắt.

Hệ thống cảnh báo của UnitedHealth đã bị cấm sử dụng bởi các chính phủ của nhiều bang sau khi bất hợp pháp hạn chế những phê duyệt điều trị sức khỏe tâm thần. PxDx của Cigna là một hệ thống cho phép bác sĩ từ chối các yêu cầu bồi thường hàng loạt chỉ bằng một cú nhấp chuột mà không cần mở hồ sơ bệnh nhân. Thuật toán DR của eviCore đã nâng tỷ lệ từ chối phê duyệt trước lên cao nhất 20 phần trăm.

TIERS ở Texas và TEDS ở Tennessee của Deloitte đã làm mất quyền đủ điều kiện Medicaid cho hàng trăm ngàn người có thu nhập thấp và người khuyết tật

do lỗi chương trình và địa chỉ gửi sai. Các thông báo đã gửi đến các ngôi nhà sai, và vì không có câu trả lời, họ được xử lý là không đủ điều kiện.

Nếu đếm các quốc gia được đề cập cho đến nay, đó là Úc, Ấn Độ, Jordan, Brazil, Anh, Đan Mạch, Thụy Điển, Pháp, Hà Lan và Hoa Kỳ. Các hệ thống khác nhau và các ngôn ngữ khác nhau.

Nhưng những người bị bắt gặp thật ngạc nhiên giống nhau. Họ là những người nghèo, già, bị bệnh, người nhập cư và không có nơi nào để nói lên tiếng nói.

Đó không phải là sự tình cờ. Các hệ thống này được giới thiệu để tiết kiệm tiền, và để tiết kiệm tiền bạn phải cắt giảm thanh toán, và để cắt giảm thanh toán bạn phải cắt loại bỏ ai đó. Đối tượng an toàn nhất để cắt loại bỏ là những người không thể phản đối.

Các thuật toán rất tốt trong tính toán đó.

1.3

Việc công nghệ nhận diện khuôn mặt nhận diện sai người da màu và bắt giữ oan

Cô gái hai tuổi và cô bé năm tuổi đang nhìn từ sân nhà.

Ngày 9 tháng 1 năm 2020, ở sân nhà trước một căn nhà ở Farmington Hills, Michigan, Robert Williams vừa tan làm về. Cảnh sát Detroit ập vào và còng tay anh trước mặt vợ con.

Cáo buộc là trộm đồng hồ. Đó là vụ thất thoát năm chiếc đồng hồ trị giá 1.800 đô-la tại cửa hàng Shinola ở Detroit vào tháng 10 năm 2018.

Bằng chứng như sau. Có cảnh quay mờ từ camera giám sát cửa hàng. Một chuyên gia pháp y của cảnh sát bang Michigan chụp một khung hình từ cảnh quay đó và đưa vào cơ sở dữ liệu nhận diện khuôn mặt. Hệ thống đưa ra ảnh bằng lái xe cũ của Williams làm ứng viên.

Phần sau đây là quan trọng. Một nhân viên bảo vệ cửa hàng, người thậm chí không có mặt tại hiện trường, nhận được nhiều bức ảnh của những người đàn ông da đen và chọn Williams.

Williams ở trong trại tạm giữ suốt ba mươi giờ.

Trong phòng thẩm vấn, một thám tử đặt ảnh từ camera giám sát lên bàn và hỏi. Người này có phải là bạn không. Williams cầm ảnh đặt bên cạnh mặt mình và nói. Người này không phải là tôi. Các bạn có nghĩ rằng tất cả người da đen đều trông giống nhau không.

Câu chuyện này có phần tiếp theo. Vào thời điểm đó, Giám đốc Cảnh sát Detroit James Craig đã biết rằng công nghệ này nhầm định nghi phạm với tỷ lệ 96 phần trăm.

Mang một công cụ sai bốn lần trong mười lần đến sân nhà của người ta.

Robert Williams trở thành hồ sơ chính thức đầu tiên ở Mỹ bị bắt giữ trái phép do lỗi nhận diện khuôn mặt. Tháng 6 năm 2024, thành phố Detroit đã nhất trí với Liên minh Tự do Dân sự Mỹ với số tiền 300.000 đô-la, và thay đổi quy định sao cho không thể đăng đơn xin lệnh bắt chỉ dựa trên manh mối nhận diện khuôn mặt và so sánh ảnh.

Tính đến năm 2026, có hơn mười hai người được biết đến là bị bắt giữ theo cách này. Đó chỉ là những con số đã biết.

Ngày 10 tháng 6 năm 2026, một vụ kiện mới cũng được nộp. Robert Dillan của Florida bị bắt năm 2024. Bằng chứng là cảnh quay giám sát mờ và kết quả độ chính xác 93 phần trăm. Trong đơn xin lệnh không có bằng chứng chứng minh rằng anh ta không thể là kẻ phạm tội.

Con số 93 phần trăm quyến rũ mọi người. Vì nó giống điểm thi thế. Nhưng con số này không có nghĩa là có 93 phần trăm khả năng người này là kẻ phạm tội. Nó có nghĩa là trong hàng triệu bức ảnh trong cơ sở dữ liệu, bức ảnh này giống nhất.

Có một người giống nhất lúc nào cũng tồn tại. Ngay cả khi kẻ phạm tội không nằm trong số đó.

Nếu không hiểu sự khác biệt này, điều gì sẽ xảy ra được thể hiện qua câu chuyện của Angela Lipps.

Ngày 14 tháng 7 năm 2025, Tennessee. Lipps là mẹ của năm đứa con và bà của năm cháu. Bà đang ở nhà chăm sóc các cháu. Các thẩm phán liên bang bước vào với súng. Người ta nói bà là kẻ trốn chạy từ một vụ lừa đảo ngân hàng xảy ra ở Fargo, North Dakota.

Lipps không bao giờ đi đến North Dakota. Bà không bao giờ lên máy bay. Bà thậm chí hiếm khi bước ra khỏi nhà quá 160 kilomet.

Những gì các thám tử Cảnh sát Fargo điều tra chỉ là điều này. Họ nhận được kết quả nhận diện khuôn mặt, tìm thấy ảnh trên phương tiện truyền thông xã hội và so sánh bằng mắt. Họ không gọi điện thoại một lần nào.

Lipps được chuyển đến North Dakota và bị giam năm một trăm lẻ tám ngày. Chỉ sau khi luật sư nộp bản ghi lịch sử ngân hàng chứng minh rằng bà ở Tennessee lúc đó, bà mới được ph釋.

Trong thời gian đó, bà đã mất nhà ở. Bà đã mất chiếc xe. Bà đã mất con chó nuôi.

Cùng một hệ thống của cảnh sát Detroit tiếp tục cắn người.

Tháng 7 năm 2019, Michael Oliver bị bắt vì cáo buộc trộm điện thoại của giáo viên. Chương trình của DataWorks Plus đã kết nối ảnh của anh trong cơ sở dữ liệu của cảnh sát với kẻ phạm tội.

Tháng 2 năm 2023, Porsha Woodruff bị bắt vì cáo buộc cướp và cướp xe. Cô ấy đang mang thai tám tháng. Cô cảm thấy chuyển dạ trên sàn trại tạm giữ.

Điều tra viên đã chạy cảnh quay giám sát qua thuật toán để lấy ảnh bằng lái xe của Woodruff từ tám năm trước. Họ không xác minh đặc điểm thân thể của mục tiêu cần tìm. Người được quay bằng camera giám sát không có thai.

Không cần trí tuệ nhân tạo để phân biệt giữa người có bụng như núi Nam Sơn và người không. Bạn cần mắt. Nhưng quy trình nhìn bằng mắt đã bị bỏ qua.

Tháng 4 năm 2025, ở New York, Travis Williams bị bắt với cáo buộc là nghi phạm phạm tội dâm ô công khai và bị giam hơn hai ngày. Lúc đó, anh ở Brooklyn, cách đó 19 kilomet.

Khi so sánh tình trạng thân thể, bạn sẽ cười. Kẻ phạm tội thực tế cao 167 centimet và nặng 72 kilogam. Williams cao 188 centimet và nặng 104 kilogam.

Cảnh sát từ chối kiểm tra dữ liệu vị trí và từ chối liên hệ với người sử dụng lao động.

Tháng 11 năm 2022, Randal Quran Reed, cư dân Georgia, bị kết tội trộm túi xách hàng hiệu ở Louisiana và bị giam sáu ngày. Lệnh bắt được cấp dựa trên một kết quả khớp từ Clearview AI. Reed nói anh ở Georgia lúc đó nhưng không ai xác minh. Sau đó, vụ kiện đã im lặng bị bác bỏ.

Tại sao lại chỉ những khuôn mặt này mới sai? Câu trả lời đã có năm 2018 rồi.

Joy Buolamwini và Timnit Gebru của MIT và Stanford đã phát hành một báo cáo có tên là Gender Shades. Đây là kết quả của việc kiểm tra các chương trình phân tích khuôn mặt của IBM, Face++, và Microsoft.

Tỷ lệ lỗi trong việc xác định giới tính của nam giới có da sáng là dưới 0,8 phần trăm. Phụ nữ da tối sai tới 34,7 phần trăm.

Đó là khoảng cứ ba cái thì sai một cái.

Nguyên nhân nằm ở dữ liệu đào tạo. Các tập dữ liệu nổi tiếng như ImageNet, LFW, và BDD100K chứa phần lớn ảnh nam giới da trắng. Máy tính phân biệt tốt khuôn mặt mà nó thấy nhiều. Những khuôn mặt nó thấy ít thì nó lại gom chúng lại.

Khi Viện Công nghệ Georgia phân tích tập dữ liệu BDD100K cho xe tự lái, cùng một mô hình cũng xuất hiện. Độ chính xác trong việc nhận ra người đi bộ da tối bình quân thấp hơn 4,8 phần trăm, tối đa thấp hơn 12 phần trăm.

Đó là nói rằng xe hơi không nhận ra mọi người. Điểm này sẽ xuất hiện lại ở phía sau.

Cảnh báo đã đủ. Bán vẫn tiếp tục.

Clearview AI đã thu thập ba tỷ bức ảnh khuôn mặt của mọi người từ Facebook, Instagram, Twitter và LinkedIn mà không có sự đồng ý để tạo ra một cơ sở dữ liệu tìm kiếm. Các công ty nền tảng đã yêu cầu ngừng lại và một vụ kiện vi phạm Đạo luật Bảo vệ Thông tin Sinh trắc học bang Illinois đã được đệ trình, nhưng công ty vẫn bán dịch vụ tìm kiếm cho các cơ quan điều tra trên toàn thế

giới.

Chúng ta không biết bức ảnh tốt nghiệp mà chúng ta đăng tải mười năm trước hiện đang nằm trong máy tính của đồn cảnh sát nào.

Amazon cũng đã cố gắng bán một sản phẩm có tên Rekognition cho cảnh sát và Cơ quan Thực thi Di trú và Hải quan. Vào tháng 7 năm 2018, Liên đoàn Tự do Dân sự Mỹ đã thử nghiệm và phát hiện ra khuôn mặt của hai mươi tám hạ nghị sĩ và thượng nghị sĩ liên bang khớp với ảnh của tội phạm. Tỷ lệ người da màu trong số các nghị sĩ bị nhận diện sai cao một cách đặc biệt.

Vào tháng 10 năm 2019, hai mươi bảy trong số một trăm tám mươi tám vận động viên thể thao chuyên nghiệp ở khu vực New England đã bị nhận diện là tội phạm.

Amazon cho biết họ đã hướng dẫn người dùng thiết lập mức độ tin cậy từ 95 phần trăm trở lên. Tại hiện trường, tiêu chuẩn đó đã không được tuân thủ. Những người đọc sách hướng dẫn luôn là thiểu số.

Quảng cáo phóng đại cũng tồn tại. Ủy ban Thương mại Liên bang đã nêu vấn đề khi công ty IntelliVision quảng cáo rằng phần mềm nhận dạng khuôn mặt của họ không có thành kiến về giới tính và chủng tộc, đồng thời có độ chính xác cao nhất thế giới. Dù tuyên bố đã được huấn luyện với hàng triệu bức ảnh, nhưng thực tế đó chỉ là một tập dữ liệu gồm một trăm nghìn bức ảnh đã được biến đổi.

Từ năm 2012 đến năm 2020, chuỗi bán lẻ Rite Aid đã tập trung lắp đặt camera nhận dạng khuôn mặt tại các cửa hàng ở những khu phố có đông người thu nhập thấp và người da màu sinh sống. Hàng nghìn khách hàng đã bị buộc tội là kẻ trộm, bị đuổi khỏi cửa hàng hoặc bị khám xét. Ủy ban Thương mại Liên bang đã cấm sử dụng công nghệ này trong năm năm.

Đến đây vẫn chỉ là câu chuyện của những người còn sống sót trở về.

Trong vụ cướp cửa hàng Sunglass Hut ở Houston, Texas vào tháng 1 năm 2022, hệ thống nhận dạng khuôn mặt của Macy's và EssilorLuxottica đã chỉ đích

danh Harvey Murphy Jr. là thủ phạm. Lúc đó ông đang ở Sacramento, California.

Bị bắt vào tháng 10 năm 2023 và bị giam giữ trong hai tuần, Murphy đã bị ba tù nhân tấn công hội đồng và tấn công tình dục. Ông bị để lại những khuyết tật thể chất vĩnh viễn.

Tại Ponddale, Missouri, đã có một vụ hành hung tại sân ga xe lửa vào năm 2021. Nghi phạm đã che mặt bằng khẩu trang và quần áo. Cảnh sát điều tra đã đưa bức ảnh đó vào thuật toán.

Ngay cả khi đưa vào một khuôn mặt bị che khuất, hệ thống vẫn đưa ra các ứng cử viên. Đó là vì nó không có chức năng trả lời là không biết. Trong số các ứng cử viên được đưa ra có Christopher Gatlin, hai mươi chín tuổi, một người cha của bốn đứa trẻ chưa từng có tiền án.

Ngay cả trong các hướng dẫn nội bộ của cảnh sát cũng ghi rõ rằng không được bắt giữ chỉ dựa vào kết quả nhận dạng khuôn mặt. Cảnh sát điều tra đã lập một hàng ảnh để nhận diện, và Gatlin đã phải ngồi tù hơn hai năm.

Kimberly Williams ở Oklahoma đã bị giam giữ sáu tháng kể từ tháng 6 năm 2021. Các cảnh sát điều tra đã sao chép kết quả tìm kiếm do một điều tra viên ngân hàng tư nhân đăng lên vào đơn xin lệnh bắt giữ mà không cần xác minh. Họ thậm chí còn không thông báo cho tòa án về việc đã sử dụng nhận dạng khuôn mặt. Bà đã mất việc làm trong khi bị chuyển qua lại giữa ba nhà tù của hạt.

Francisco Arteaga ở New Jersey đã bị bắt vào tháng 11 năm 2019 vì tình nghi cướp có vũ trang và phải ngồi tù chờ xét xử trong bốn năm. Đó là do các cơ quan điều tra đã từ chối tiết lộ nguyên lý hoạt động, tỷ lệ lỗi và mã nguồn của hệ thống. Khi không biết điều gì đã chỉ điểm mình, người ta sẽ không có cách nào để phản bác.

Alonzo Sawyer, năm mươi tư tuổi, ở Baltimore, Maryland, đã bị giam giữ chín ngày với tư cách là nghi phạm trong một vụ hành hung trên xe buýt. Ông có chiều cao khác, tuổi tác khác, khoảng trống giữa hai răng cửa khác và dáng đi khác với nghi phạm thực sự.

Những chuyện tương tự cũng xảy ra bên ngoài biên giới.

Vào tháng 1 năm 2026 tại Southampton, Vương quốc Anh, kỹ sư phần mềm gốc Bangladesh tên là Albi Chowdhury đã bị bắt giữ ngay trước mặt cha mẹ và hàng xóm. Anh đã bị chỉ định là nghi phạm trong một vụ trộm nhà hoang ở Milton Keynes cách đó một trăm tám mươi lăm km. Anh đã bị tạm giam trong mười giờ.

Hệ thống do Cảnh sát Thames Valley sử dụng là sản phẩm của công ty Cognitec của Đức, và nó chứa một thuật toán cũ được phát hành vào năm 2020. Theo đánh giá của Phòng thí nghiệm Vật lý Quốc gia Vương quốc Anh, trong một số cài đặt nhất định, tỷ lệ nhận dạng sai đối với người châu Á là 4.0 phần trăm. Đối với người da trắng là 0.04 phần trăm.

Gấp một trăm lần.

Các sĩ quan cảnh sát khẳng định rằng họ đã kiểm tra lại bằng mắt người. Hiện tượng này được gọi là thiên kiến tự động hóa. Khi máy móc đưa ra câu trả lời, con người thay vì xác minh câu trả lời đó thì lại cố gắng làm cho phù hợp với nó. Họ bắt đầu tìm kiếm những điểm giống nhau. Việc tìm kiếm những điểm giống nhau trên khuôn mặt con người là điều không khó.

Vào tháng 5 năm 2024 tại nhà ga London Bridge, Shaun Thompson, ba mươi tám tuổi, đã bị bắt giữ. Anh là một nhà hoạt động vì quyền lợi thanh thiếu niên đang trên đường về nhà sau khi hoàn thành công việc tình nguyện phòng chống tội phạm dùng dao. Hệ thống nhận dạng khuôn mặt theo thời gian thực của Cảnh sát Metropolitan đã chỉ định anh là một tội phạm đang bị truy nã, và anh đã phải nhận yêu cầu xuất trình giấy tờ tùy thân cùng với lời đe dọa bắt giữ trong ba mươi phút.

Hệ thống kiểm tra ảnh hộ chiếu của Bộ Nội vụ Vương quốc Anh đã đọc sai môi của những người nộp đơn có màu da sẫm thành trạng thái đang mở miệng. Đơn xin cấp hộ chiếu của những người nộp đơn da đen như Joris Lesein và Joshua Bada đã bị từ chối. Bộ Nội vụ đã đưa hệ thống vào áp dụng mặc dù biết tỷ lệ thất bại cao đối với người da màu.

Lực lượng Phòng vệ Israel đã triển khai các hệ thống mang tên Red Wolf, Blue Wolf và Corsight tại các trạm kiểm soát ở Dải Gaza và Bờ Tây. Họ đã thu thập khuôn mặt của cư dân Palestine mà không có sự đồng ý để thiết lập một mạng lưới giám sát tự động. Do lỗi, nhiều cư dân vô tội đã bị coi là nghi phạm khủng bố, bị bắt đi, thẩm vấn và đánh đập.

Tại Hyderabad, Ấn Độ, vào tháng 2 năm 2023, một lao động công nhật ba mươi lăm tuổi tên là Mohammed Kadir đã bị bắt giải đi. Lý do là vì khuôn mặt của anh giống với người trong đoạn video về một vụ móc túi. Nhân vật trong đoạn video đó đang dùng khăn che mặt.

Kadir đã bị giam giữ và tra tấn trong năm ngày rồi qua đời.

Tại Brazil vào tháng 4 năm 2024, João Antônio đang đi xem một trận bóng đá thì bị còng tay vì khuôn mặt của anh giống với một tội phạm đang bị truy nã. Hệ thống nhận dạng người bị truy nã ở Buenos Aires, Argentina đã bị đình chỉ hoạt động sau khi gây ra hơn một trăm bốn mươi vụ bắt giữ nhầm.

Khi đặt những sự kiện này cạnh nhau, một trình tự sẽ xuất hiện.

Các công ty công nghệ phóng đại độ chính xác để bán hàng. Cảnh sát sử dụng kết quả đó như vật chứng. Trong đơn xin lệnh bắt giữ, cụm từ sử dụng nhận dạng khuôn mặt bị bỏ qua. Thẩm phán ký tên mà không hay biết. Người bị bắt không biết tại sao mình lại bị chỉ định. Khi bị hỏi về nguyên lý hoạt động, họ nói đó là bí mật kinh doanh.

Gánh nặng chứng minh đã bị đảo ngược. Thay vì nhà nước chứng minh tội lỗi, công dân phải chứng minh vô tội. Để làm được như vậy, họ cần bằng chứng và

tiền luật sư.

Angela Lipps mất 108 ngày. Christopher Gatlin mất 2 năm. Francisco Arteaga mất 4 năm.

Mohamed Kadir không còn thời gian.

CHƯƠNG 2

Ảo giác và thông tin sai lệch: Khủng hoảng niềm tin của AI tạo sinh

2.1

Ảo giác của AI tạo sinh và sự phát tán sách, bài báo giả mạo

Bất kỳ tác giả nào thỉnh thoảng cũng thử tìm kiếm tên của chính mình. Đó là một thói quen đáng xấu hổ, nhưng ai cũng làm vậy.

Vào một buổi sáng tháng 8 năm 2023, nữ tác giả đồng thời là nhà phê bình sách người Mỹ Jane Friedman cũng làm như vậy. Danh sách các cuốn sách của bà hiện lên trên màn hình. Tuy nhiên, có năm cuốn trông rất xa lạ.

Đó là những cuốn sách bà chưa từng đọc. Đó cũng là những cuốn sách bà chưa từng viết.

Những cuốn sách mang tên Friedman nằm ở đầu kết quả tìm kiếm của Amazon và Goodreads đó là sản phẩm do trí tuệ nhân tạo tạo sinh tạo ra. Nếu mở ra xem, các câu văn có phần trơn tru nhưng không có cốt lõi. Đó là đặc điểm của các bài viết do mô hình ngôn ngữ lớn tạo ra. Không có nhiều từ sai, nhưng cũng không đọng lại được gì nhiều.

Danh tiếng cả đời tích lũy đã bị lợi dụng bởi những văn bản kém chất lượng chỉ trong một đêm.

Friedman đã yêu cầu Amazon xóa bỏ chúng. Một lá thư trả lời đã được gửi đến. Quý khách đã không đăng ký tên của mình dưới dạng nhãn hiệu.

Điều này có nghĩa là để sử dụng tên của chính mình, bà đã phải đăng ký nhãn hiệu trước. Đây là một trong những câu nói mang tính quan liêu bậc nhất trên thế giới.

Chỉ sau khi Friedman đăng tải sự việc này lên blog và mọi người tỏ ra tức giận, Amazon mới lặng lẽ gỡ bỏ những cuốn sách đó. Không phải quy định đã thay đổi, mà họ gỡ xuống vì sự việc trở nên ồn ào.

Nếu những cuốn sách giả mạo chỉ dừng lại ở việc đánh cắp tên tuổi thì đã là điều may mắn, nhưng mọi chuyện không dừng lại ở đó.

Vào tháng 9 năm 2023, Hiệp hội Nấm học New York đã đưa ra một cảnh báo khẩn cấp. Nội dung là nhiều cuốn cẩm nang hướng dẫn hái nấm dại dành cho người mới bắt đầu được bán trên Amazon là sách do ChatGPT viết.

Cốt lõi của cuốn sách về nấm chỉ có một. Đó là cách phân biệt nấm có thể ăn được và nấm ăn vào sẽ tử vong. Phần đó đã bị sai.

Đó không phải là lỗi đánh máy. Đó là thông tin có thể khiến con người tử vong.

Vào tháng 2 năm 2024, ngay khi tin tức về việc Vua Charles III được chẩn đoán mắc bệnh ung thư được đưa ra, bảy cuốn tiểu sử giả mạo đã xuất hiện trên Amazon. Nội dung cho rằng Nhà vua mắc bệnh ung thư da hoặc bị tai nạn bị che giấu được trộn lẫn giống như tin tức có thật. Cung điện Buckingham đã rất tức giận.

Viết nhanh, đăng tải nhanh và chiếm lĩnh vị trí đầu trong kết quả tìm kiếm để kiếm tiền. Trước khi trí tuệ nhân tạo xuất hiện, đây là công việc mà con người phải mất vài ngày để làm. Giờ đây chỉ mất vài phút.

Thỉnh thoảng, sự tự động hóa này lại tự bộc lộ chân tướng.

Vào tháng 1 năm 2024, những sản phẩm kỳ lạ đã xuất hiện trên Amazon. Đó là ghế sân vườn, vòi nước và bàn. Tại vị trí tên sản phẩm, có ghi dòng chữ như thế này.

Xin lỗi. Yêu cầu này vi phạm chính sách điều khoản sử dụng của OpenAI nên không thể thực hiện được.

Người bán đã yêu cầu trí tuệ nhân tạo viết mô tả sản phẩm, trí tuệ nhân tạo đã từ chối và họ đã sao chép y nguyên câu từ chối đó rồi đăng lên. Thậm chí họ còn không đọc qua.

Đó là một tình huống buồn cười. Tuy nhiên, nếu nhìn lại điều này thì thật đáng sợ. Người bán này đã không đọc mô tả của món đồ mình bán. Có bao nhiêu đoạn mô tả chưa được đọc đã được đăng tải lên Internet?

Vào tháng 4 năm 2024, Google Books đã thu thập nguyên xi những cuốn sách chất lượng thấp do trí tuệ nhân tạo viết vào cơ sở dữ liệu tìm kiếm sách. Trong số đó, nhiều cuốn vẫn còn lại câu nói này. Theo bản cập nhật kiến thức mới nhất của tôi.

Có một công cụ mang tên Google Ngram Viewer. Đây là một công cụ học thuật theo dõi dữ liệu sách xem một từ vựng nào đó đã được sử dụng bao nhiêu lần vào thời đại nào. Đó là thước đo lịch sử của ngôn ngữ. Nếu những cuốn sách do trí tuệ nhân tạo viết bị trộn lẫn vào thước đo đó, vạch chia sẽ bị mờ đi.

Đến đây là câu chuyện của ngành xuất bản. Các tòa soạn báo còn vấp ngã đau hơn.

Vào tháng 5 năm 2025, tờ Chicago Sun-Times đã xuất bản một phụ bản đặc biệt mang tên danh sách sách gợi ý cho mùa hè. Các tác phẩm mới của các tác giả nổi tiếng như Isabel Allende, Andy Weir và Lee Min-jin đã được đăng tải cùng với những lời nhận xét giới thiệu hào nhoáng.

Độc giả đã đến hiệu sách. Sách không có ở đó.

Những cuốn sách có trong danh sách đó không một cuốn nào tồn tại trên thế giới. Tác giả là nhân vật có thật nhưng chỉ có tiêu đề sách là do trí tuệ nhân tạo tự bịa ra. Cách thức gắn thông tin giả vào tên thật là hình thức nói dối mà trí tuệ nhân tạo làm giỏi nhất. Nếu tìm kiếm để xác minh, tên tác giả là thật nên người ta sẽ cảm thấy yên tâm.

Tòa soạn báo đã cắt giảm nhân sự và nhận nội dung từ các công ty bên ngoài để sử dụng. Một nhà văn tự do đã sao chép y nguyên danh sách mà trí tuệ nhân tạo nhả ra, và ban biên tập đã không kiểm tra một lần nào.

Đó là khoảnh khắc cái tên của một tờ báo lớn đã đóng dấu tín nhiệm cho lời nói dối của trí tuệ nhân tạo.

Vào tháng 11 năm 2023, Sports Illustrated đã bị phát hiện. Trang tin tức điều tra Futurism đã phơi bày việc phương tiện truyền thông này liên tục đăng tải các bài báo do trí tuệ nhân tạo viết.

Tên phóng viên, ảnh khuôn mặt và kinh nghiệm làm việc ghi trên bài báo đều là giả. Ảnh khuôn mặt đã được tạo ra bằng thuật toán tổng hợp. Một phóng viên không tồn tại với kinh nghiệm làm việc không tồn tại đã viết bài đánh giá sản phẩm.

Phương tiện truyền thông đã đổ lỗi cho công ty gia công bên ngoài. Việc cắt giảm các phóng viên là con người và áp dụng tự động hóa là quyết định do ban giám đốc đưa ra.

Tại một tờ báo địa phương, một phóng viên đã trực tiếp nhúng tay vào.

Vào tháng 8 năm 2024, Aaron Pelczar, một phóng viên mới của tờ Cody Enterprise ở bang Wyoming, đã từ chức. Phóng viên của một tờ báo đối thủ đã phát hiện ra những câu văn trơn tru một cách kỳ lạ trong bài báo của anh ta và đã gọi điện trực tiếp cho nguồn tin, từ đó sự việc bị lộ tẩy.

Pelczar đã bịa ra phát ngôn của sáu quan chức, bao gồm cả thống đốc bang, bằng ChatGPT rồi đăng lên bài báo. Anh ta đã biến những lời chưa từng nói thành những lời đã nói.

Phần ngoạn mục nhất lại nằm ở chỗ khác. Cuối bài báo anh ta viết, dòng câu hướng dẫn của trí tuệ nhân tạo giải thích về kỹ thuật viết tin tức hình tam giác ngược vẫn còn sót lại mà chưa bị xóa. Điều này chẳng khác nào việc dán tờ giấy quay cóp vào bài thi rồi nộp.

Vào tháng 5 năm 2023, tờ Irish Times đã đăng tải một bài viết ý kiến cho rằng việc nhuộm da nâu nhân tạo mang tính phân biệt chủng tộc, nhưng sau đó rút lại và xin lỗi khi phát hiện ra toàn bộ bài báo đã bị thao túng bởi trí tuệ nhân

tạo.

Bây giờ chúng ta sẽ đi đến phần thực sự ớn lạnh. Trí tuệ nhân tạo tạo ra những sự kiện không tồn tại.

Vào tháng 12 năm 2023, ứng dụng tin tức NewsBreak đã phát hành một bài báo nói rằng một người phụ nữ đã bị bắn chết ở Bridgeton, bang New Jersey vào đúng ngày Giáng sinh. Người dân địa phương đã hoảng hốt và cảnh sát đã vào cuộc xác minh.

Không hề có sự việc nào như vậy.

Công cụ viết lại tin tức mà NewsBreak đang sử dụng đã nhào nặn những thông tin rời rạc trôi nổi trên Internet và tạo ra một vụ giết người.

Vào tháng 10 năm 2024, mạng tin tức địa phương Hoodline đã phát hành một bài báo nói rằng Ủy viên Công tố Hạt San Mateo bang California, John Casciano Thompson, đã bị truy tố với tội danh giết người. Người khởi tố tội phạm lại trở thành kẻ sát nhân.

Điều tương tự cũng xảy ra đối với các cá nhân.

Vào tháng 6 năm 2023, người dẫn chương trình radio Mark Walters đã kiện OpenAI về phỉ báng sau khi phát hiện ChatGPT tạo ra một bản tóm tắt phiên tòa giả mạo mô tả ông là một tên lừa đảo đã chiếm dụng hơn năm triệu đô la.

Đây là cách nó xảy ra. Nhà báo Fred Riehl yêu cầu ChatGPT tóm tắt nội dung của một vụ kiện thực sự của một tổ chức bảo vệ quyền súng. ChatGPT đã đưa Walters xuất hiện, không liên quan gì đến vụ kiện đó, và biến ông thành kẻ chiếm dụng. Với các con số vụ án và câu văn có vẻ thuyết phục.

Công ty dịch vụ bản đồ OpenCage đã phải nhận những cuộc gọi phản đối vô cớ vì ChatGPT nói dối rằng công ty họ cung cấp dịch vụ tra cứu số điện thoại.

Vào tháng 3 năm 2025, Arve Jarmar Holm người Na Uy đã hỏi ChatGPT về tên của anh. Câu trả lời trở lại như thế này. Anh ấy đã giết hai đứa con và cố gắng

giết đứa con thứ ba trước khi bị kết án 21 năm tù.

Trong câu này, quê hương và số lượng con em là thực tế. Phần còn lại là giả tạo.

Cách trộn lẫn sự thật và giả tạo này là thói quen của máy tính xác suất. Nó không lưu trữ và truy xuất các sự kiện, mà thay vào đó chọn những từ có khả năng cao sẽ theo sau những từ trước đó. Khi chọn câu văn có vẻ tự nhiên theo sau tên của người Na Uy, nó sẽ tạo ra các bài báo về các sự kiện. Máy tính không biết rằng đó là cuộc sống của một con người.

Khi chính trị bị liên quan, thói quen này trở thành một vũ khí.

Vào tháng 11 năm 2023, trang tin tức Pakistan Global Village Space đã xuất bản một bài báo rằng Moshe Yatom, bác sỹ tâm thần riêng của Thủ tướng Israel Benjamin Netanyahu, đã tự tử sau khi để lại lá thư từ biệt chỉ trích tính tàn ác của Thủ tướng.

Không có người nào tên Moshe Yatom. Không có lá thư từ biệt. Tất cả đều được bịa đặt. Bài báo này đã được dịch sang nhiều ngôn ngữ và lan truyền vào thời điểm cuộc chiến giữa Israel và Hamas leo thang.

Vào tháng 4 năm 2024, Grok được gắn vào X đã thu thập những tin đồn đang lưu hành trên internet và tạo ra tiêu đề cho rằng Iran đã bắn tên lửa vào Tel Aviv. Nó cũng tạo ra một bài báo rằng cầu thủ bóng rổ Clay Thompson đã ném gạch trong thành phố. Cái sau dường như là kết quả của việc đọc theo nghĩa đen một thành ngữ bóng rổ có nghĩa là anh ấy đã bỏ lỡ nhiều cú sút.

Một máy tính không thể hiểu được những trò đùa đang viết tin tức.

Cách sử dụng trái phép tên của các phương tiện truyền thông là tinh vi hơn.

Ban biên tập của The Guardian ở Anh liên tục nhận được những câu hỏi kỳ lạ từ độc giả. Họ hỏi nơi nào có thể đọc những bài báo như vậy. The Guardian không bao giờ viết những bài báo đó.

Khi yêu cầu ChatGPT cung cấp tài liệu, nó đã kết hợp tên của các nhà báo thực sự có trên The Guardian với những tiêu đề có vẻ thuyết phục và năm xuất bản. Nó đã sử dụng các phương tiện truyền thông thực tế và các nhà báo thực tế để trao quyền hạn cho những lời nói dối. Chris Moran, biên tập viên chịu trách nhiệm về đổi mới tại The Guardian, đã công khai tố cáo hiện tượng này.

Vào tháng 7 năm 2024, Phòng thí nghiệm Báo chí Nieman đã thực hiện một thí nghiệm thích hợp về vấn đề này. Họ yêu cầu ChatGPT cung cấp các liên kết đến các bài báo điều tra từ AP, Wall Street Journal, Financial Times, The Times, Boston Globe, và Politico. Đây là những phương tiện truyền thông đã ký một hợp đồng dữ liệu chính thức với OpenAI.

ChatGPT đã cung cấp các địa chỉ trông giống như thật cùng với những bản tóm tắt chi tiết. Tất cả các địa chỉ đó đều không tồn tại.

Giới học thuật cũng giống nhau. Các bài viết không tồn tại của Giáo sư Henrik Enghoff đã được trích dẫn và một sự kiện lịch sử không tồn tại gọi là Holocaust do chết đuối đã được tạo ra.

Trí tuệ nhân tạo được gắn vào thanh tìm kiếm đã mang vấn đề này vào nhà.

Google AI Overview đã từng trả lời rằng để ngăn phô mai pizza không chảy, nên trộn một chút keo vào. Đây là kết quả của việc đọc nghiêm túc một trò đùa trên diễn đàn internet. Nó cũng đã đặt hình ảnh của một loại nấm độc ở đầu các đề xuất về nấm ăn được.

Một nghiên cứu học thuật đã báo cáo rằng 43 phần trăm các thuật ngữ tìm kiếm liên quan đến tài chính có lỗi nghiêm trọng. Google đã tiếp tục dịch vụ. Vì nó không thể từ bỏ thị trường tìm kiếm.

Có các sự kiện ở Hàn Quốc. Trí tuệ nhân tạo sinh thành của Naver đã tóm tắt về Dokdo là một vùng tranh chấp và lãnh thổ của Nhật Bản dựa trên các tuyên bố của Bộ Ngoại giao Nhật Bản để trả lời câu hỏi liên quan đến Dokdo.

Thông tin giả cũng có thể bước ra khỏi màn hình.

Vào tháng 10 năm 2024, những người đó đã nườn nhạp vào trung tâm Dublin ở Ireland. Đó là do thông tin về một cuộc diễu hành Halloween không tồn tại do ChatGPT tạo ra đã xuất hiện ở đầu kết quả tìm kiếm. Hàng nghìn công dân và du khách đã đứng trên những con đường mưa chờ đợi một cuộc diễu hành sẽ không bao giờ xảy ra.

Vào tháng 11 năm 2025, tại London, các bức ảnh AI và bài báo nói rằng một thị trường Giáng sinh hoàng gia lộng lẫy được mở trước Cung điện Buckingham đã lan truyền trên phương tiện truyền thông xã hội. Các du khách từ khắp nơi trên thế giới đã đứng trước những cánh cổng sắt đóng cửa và vũng nước mưa.

Vào tháng 1 năm 2026, tại Tasmania, Úc, các du khách đã tin tưởng vào một bài báo được viết bởi trí tuệ nhân tạo và tìm kiếm một suối nước nóng không tồn tại, cõng lạc trong vùng đất hoang.

Gốc rễ của tất cả các sự kiện này nằm ở bản chất của công nghệ.

Mô hình ngôn ngữ lớn không phải là một máy tính biết các sự kiện. Nó là một máy chọn những từ có khả năng cao sẽ theo sau những từ trước đó. Nó không có một thiết bị bên trong để phán đoán những gì là sự thật. Việc không có thiết bị đó không phải là một lỗi, mà là thiết kế.

Vậy thì việc làm cho mô hình lớn hơn có cải thiện không? Các nghiên cứu công bố năm 2024 đã cho kết quả ngược lại. Khi mô hình trở nên lớn hơn, thay vì nói rằng nó không biết, nó có xu hướng đưa ra những câu trả lời sai lạc một cách tự nhiên hơn và tự tin hơn.

Nó không trở nên thông minh hơn mà là khả năng nói chuyện của nó tăng lên.

Các biên tập viên và những người kiểm tra sự kiện là những người được cho là phải ngăn chặn vấn đề này. Tuy nhiên, những vị trí đó đã được phân loại như chi phí và biến mất trước tiên.

Khi sự cố xảy ra, các công ty đưa ra cùng một câu trả lời. Nó vẫn là một dịch vụ beta. Chúng tôi đã viết ở dưới cùng của màn hình rằng nó có thể không chính

xác.

Những chữ nhỏ đó không đến được với người mua sách về nấm.

2.2

Vụ bê bối về án lệ và trích dẫn giả mạo trong tòa án và giới học thuật

Nó bắt đầu với một cái đầu gối.

Roberto Mata đã va chạm đầu gối vào xe đẩy trên máy bay của Avianca. Ông ấy đã kiện hãng hàng không để đòi bồi thường. Đó là một vụ việc thường gặp. Những vụ kiện như vậy không được ghi vào bộ sưu tập các phán quyết.

Vào tháng 5 năm 2023, vụ việc này đã trở thành một vụ việc mà cộng đồng pháp lý toàn thế giới biết đến. Không phải vì cái đầu gối.

Luật sư của Mata, Stephen Schwartz, đã đưa ra sáu phán lệ để chống lại đơn yêu cầu bác bỏ của hãng hàng không. Chúng là những tên gọi giống như vụ Vargas so với China Southern Airlines. Chúng có số vụ, ngày phán quyết, và tên tòa án. Hình thức là hoàn hảo.

Đội luật sư của bên đối phương đã cố gắng tìm các phán lệ này. Chúng không tồn tại. Các phụ tá của thẩm phán đã tìm. Chúng không tồn tại.

Tất cả sáu phán lệ đều là những phán quyết không tồn tại.

Schwartz đã sử dụng ChatGPT để chuẩn bị tài liệu. Trong phiên thẩm vấn, ông nói rằng ông không biết rằng trí tuệ nhân tạo có thể nói dối. Ông thậm chí còn hỏi ChatGPT xem các phán lệ này có thật không. ChatGPT đã trả lời rằng chúng là thật.

Đó là như hỏi một kẻ nói dối liệu nó có nói dối không.

Luật sư đồng sự Peter Loduca đã ký mà không xác nhận. Tòa án đã phạt cả hai người năm ngàn đô la.

Lý do vụ việc này trở nên nổi tiếng không phải vì số tiền phạt. Đó là vì lần đầu tiên, mọi người nhìn sâu vào cách các tài liệu được nộp cho tòa án được tạo ra.

Và điều này đã lặp lại. Nó đã lặp lại nhiều lần.

Nhà nghiên cứu Pháp Damien Charlotin đã tạo ra một cơ sở dữ liệu thu thập các phán lệ ảo từ trí tuệ nhân tạo. Hãy xem những con số này tăng lên như thế nào.

Vào giữa năm 2025, nó khoảng 200 vụ. Vào tháng 1 năm 2026, nó trở thành 719 vụ. Đầu tháng 4 năm 2026 là 1.227 vụ, ngày 9 tháng 6 là 1.598 vụ.

Từ ngày 22 tháng 5 đến ngày 9 tháng 6, trong mười tám ngày, tăng 140 vụ. Tức là trung bình tám vụ mỗi ngày.

Điều này có nghĩa là các phán lệ giả được phát hiện trong tòa án xảy ra tám lần mỗi ngày. Những cái không được phát hiện không được tính.

Nhìn vào các vụ việc trong danh sách này, quốc tịch và chức vụ là đa dạng.

Michael Cohen, người gần gũi với cựu Tổng thống Donald Trump, đã sử dụng Google Bard khi chuẩn bị tài liệu yêu cầu kết thúc sớm thời gian giám sát. Nghĩ rằng các trích dẫn là thật, ông đã chuyển chúng cho luật sư của mình, và luật sư nộp chúng cho tòa án mà không xác nhận. Ông chỉ biết khi thẩm phán cho biết chúng không có trong bộ sưu tập các phán quyết.

Luật sư Jae-i Lee ở New York đã tạo các phán lệ giả bằng ChatGPT khẳng định rằng bác sĩ ở Queens đã thực hiện phẫu thuật nạo thai sai, nộp cho Tòa phúc thẩm Liên bang Cấp 2, và sau đó bị chuyển vào quá trình kỷ luật.

Tại Tòa án Tối cao của tỉnh British Columbia, Canada, hai phán lệ được trình bày là tiền lệ rằng cha có thể đưa con sang Trung Quốc trong vụ ly hôn đã bị phát hiện là ảo tưởng. Thẩm phán đã bắt luật sư đó phải trả chi phí luật sư của bên đối phương bằng tiền cá nhân.

Vụ Felicity Harber tại Tòa án Bậc nhất của Anh giống như một tiểu thuyết trinh thám. Ông không chấp nhận quyết định phạt 3.265 bảng và đã nộp chín phán lệ. Thẩm phán Ann Redstone đã phát hiện ra điều gì đó kỳ lạ. Dù là phán quyết của Anh, nhưng có chứa chính tả theo phong cách Mỹ. Và các câu cách điệu được lặp lại ở một số chỗ.

Máy không thể ẩn quốc tịch của nó.

Tòa án Tối cao của Quần đảo Canary, Tây Ban Nha đã phạt một luật sư đã nộp tới 48 phán lệ giả 420 euro. Luật sư đó thậm chí không kiểm tra với cơ sở dữ liệu pháp lý chính thức.

Tại Tòa án Gia đình Melbourne, Úc, một công cụ của phần mềm chuyên ngành pháp luật đã tạo ra các phán lệ giả, khiến phiên tòa bị hoãn lại. Tại Tòa án Tối cao Tasmania, trong một vụ kháng cáo về phỉ báng, Chủ tịch Tòa án Tối cao đã trực tiếp chỉ ra các phán lệ không tồn tại và bác bỏ vụ việc.

Ở Brazil, một thẩm phán liên bang vì gánh nặng viết phán quyết đã cho phép trợ lý sử dụng công cụ học máy, nhưng sau đó bị điều tra vì đã nhập các phán lệ của tòa án cấp cao bị biến dạng.

Những phán quyết đã được viết bởi trí tuệ nhân tạo. Người bị xét xử không biết điều này.

Vào năm 2026, quy mô đã trở nên lớn hơn.

Trong vụ kháng cáo ly hôn tháng 2 năm 2026, tài liệu được nộp bởi luật sư Greg Lake có vấn đề với 57 trích dẫn trong số 63 trích dẫn. Có 20 phán lệ giả, 3 quyết định bị thao túng, và các trích dẫn luật pháp được bịa ra.

Chỉ 6 trong 63 là thật.

Tòa án Tối cao Nebraska vào tháng 4 năm 2026 đã tạm ngưng bằng cấp của ông ấy và lệnh tiến hành điều tra kỷ luật.

Phán quyết Whiting của Tòa phúc thẩm Liên bang Cấp 6 vào tháng 3 năm 2026 đã xử lý hơn 24 trích dẫn giả và sai lệch sự thật trong ba tài liệu kháng cáo được hợp nhất. Trong một vụ tại Tòa án Liên bang Hạ cấp Oregon, tổng cộng khoảng 109.700 đô la bị áp dụng cho các biện pháp trừng phạt và chi phí. Chỉ trong quý 1 năm 2026, các biện pháp trừng phạt mà các tòa án Mỹ áp dụng cho tài liệu liên quan đến trí tuệ nhân tạo đã vượt quá 145.000 đô la.

Nhìn vào hồ sơ, một mô hình xuất hiện. Những ai cố gắng che giấu các trích dẫn giả bị phạt nặng hơn chính những trích dẫn giả. Các thẩm phán tha thứ sai lầm nhưng không tha thứ nói dối.

Các công ty cũng không phải là ngoại lệ.

DoNotPay, được quảng cáo là luật sư robot đầu tiên thế giới, đã tuyên bố có thể phân tích các tài liệu pháp lý mà không xác nhận luật sư, bị bắt bởi Ủy ban Thương mại Liên bang và phạt 193.000 đô la.

EvenUp, được cho là tự động hóa chuẩn bị tài liệu thương tích cá nhân, đã bị phơi bày những sự thật thông qua lời tiết lộ của cựu nhân viên. Khi trí tuệ nhân tạo bỏ sót chi tiết chấn thương hoặc tạo ra bệnh giả, các nhân viên phải sửa chữa bằng tay.

Trong vụ kiện bồi thường cháy hoverboard, tám phán lệ giả được tạo bởi cơ sở dữ liệu của chính công ty luật đã được nộp. Thẩm phán phạt các luật sư tổng cộng năm ngàn đô la và loại bỏ luật sư chủ đạo khỏi vụ việc.

Các công ty luật lớn cũng bị buộc tội. Trong vụ kiện công ty bảo hiểm State Farm, các luật sư của hai công ty luật khổng lồ đã sử dụng hỗn hợp Gemini và công cụ chuyên ngành pháp luật, phát hiện 9 trong 27 trích dẫn có lỗi, hai cái là hoàn toàn giả, và bị phạt 31.000 đô la.

Đội luật sư của Chủ tịch MyPillow Mike Lindell đã để lại hơn 30 phán lệ giả và bị phạt ba ngàn đô la mỗi luật sư.

Trường hợp kỳ lạ nhất đến từ bên trong một công ty trí tuệ nhân tạo.

Trong vụ kiện bản quyền giữa Universal Music Group và Anthropic, tài liệu do nhà khoa học dữ liệu của Anthropic nộp chứa các bài báo học thuật giả được tạo bởi chatbot của công ty. Công ty trí tuệ nhân tạo đã bị mắng trong tòa án vì lý do trí tuệ nhân tạo.

Giáo sư Jeff Hancock, chuyên gia thông tin truyền thông của Đại học Stanford, đã trích dẫn hai bài báo giả mạo do ảo giác tạo ra trong một bản khai nộp cho phiên tòa về Đạo luật Cấm Deepfake bang Minnesota. Một chuyên gia ra tòa để giải thích về sự nguy hiểm của deepfake lại trích dẫn tài liệu do trí tuệ nhân tạo bịa ra.

Bước ra khỏi phòng xử án là trường học. Tình hình ở đây cũng không có gì khác biệt.

Vào tháng 8 năm 2023, giáo sư Henrik Enghoff, một nhà sinh học người Đan Mạch, không thể tin vào mắt mình khi đang đọc bài báo về loài cuốn chiếu của một nhóm nghiên cứu người Ethiopia. Hai bài báo mà ông không hề viết lại được trích dẫn dưới tên của ông.

Nhóm nghiên cứu cho biết họ đã sử dụng ChatGPT để viết bài báo và chỉ sau đó mới biết rằng các trích dẫn giả đã được tạo ra. Bài báo đó đã bị thu hồi.

Trích dẫn là gia phả trong học thuật. Đó là việc ghi lại xem một ý tưởng bắt bắt nguồn từ đâu. Trích dẫn giả cũng giống như việc ghi một người không tồn tại vào gia phả. Một khi đã bị lẫn lộn, những người đi sau sẽ liên tục bước hụt.

Vấn đề này còn lâu đời hơn cả trí tuệ nhân tạo.

Từ năm 2008 đến năm 2013, nhà nghiên cứu người Pháp Cyril Labbe đã tìm thấy hơn một trăm hai mươi bài báo giả trong các kỷ yếu hội thảo học thuật của Springer và IEEE, những nhà xuất bản học thuật nổi tiếng. Những bài báo này là sản phẩm của một chương trình tự động tạo bài báo do các nghiên cứu sinh của MIT tạo ra vào năm 2005 nhằm trêu chọc các hội nghị có khâu thẩm định lỏng lẻo.

Những tổ hợp từ vô nghĩa đã vượt qua quá trình thẩm định và được đăng trên các tạp chí học thuật chính thức. Đó là chuyện của hai mươi năm trước.

Vào tháng 2 năm 2024, một tạp chí sinh học tế bào đã thu hồi một bài báo. Trong bài báo đó có những bức ảnh được tạo ra bằng Midjourney, trong đó có hình ảnh một con chuột khổng lồ không hợp lý về mặt giải phẫu học. Các chữ cái viết trên hình cũng chỉ là những dãy bảng chữ cái vô nghĩa.

Nhiều giám khảo đã xem và phê duyệt bức hình này.

Trong một cuộc điều tra của 404 Media vào năm 2024, hơn một trăm mười lăm bài báo đăng ký trên Google Scholar vẫn còn chứa câu "theo bản cập nhật kiến thức mới nhất của tôi". Họ đã dán vào mà không hề đọc lại.

Tại Journal of Human Evolution, nhà xuất bản đã đưa trí tuệ nhân tạo vào quá trình biên tập bản thảo mà không bàn bạc với ban biên tập. Khi bản thảo được thông qua bị hỏng định dạng và nội dung bị bóp méo, toàn bộ ban biên tập đã từ chức.

Nhóm nghiên cứu của Đại học Stanford tiết lộ rằng có tới 17 phần trăm các đánh giá bình duyệt của một hội nghị trí tuệ nhân tạo toàn cầu đã được viết bằng ChatGPT.

Đó là một cấu trúc nơi trí tuệ nhân tạo thẩm định các bài báo về trí tuệ nhân tạo. Người ta có thể hỏi con người đang ở đâu.

Điều tương tự cũng đã xảy ra trong các tài liệu chính sách.

Giáo sư danh dự James Guthrie của Đại học Macquarie ở Úc đã sử dụng Google Bard để viết một tài liệu trình quốc hội yêu cầu quy định ngành tư vấn. Bard đã bịa ra những sự việc như Deloitte bị kiện vì lơ là trong việc kiểm toán một công ty xây dựng yếu kém, hay KPMG tham gia vào một vụ bóc lột tiền lương tại cửa hàng tiện lợi. Tất cả đều là những việc không hề tồn tại. Giáo sư đã gửi thư xin lỗi tới thượng viện.

Tài liệu nhằm chỉ trích ngành tư vấn lại vu khống các công ty tư vấn bằng những tội danh không có thực.

Pháp nhân Deloitte tại Úc, nơi viết báo cáo về hệ thống phúc lợi của chính phủ Úc, cũng đã đăng nguyên văn các trích dẫn học thuật giả và các đoạn trích từ phán quyết của tòa án liên bang do ChatGPT tạo ra, dẫn đến việc bị chỉ trích và phải đền bù tiền.

Một báo cáo y tế do Bộ Y tế và Dịch vụ Nhân sinh Hoa Kỳ và Nhà Trắng công bố vào năm 2025 có chứa hơn bảy nghiên cứu học thuật giả mạo. Cách ghi trích dẫn đặc trưng của ChatGPT vẫn còn nguyên vẹn.

Những người phủ nhận biến đổi khí hậu đã tạo ra một bài báo bằng mô hình Grok 3 và đăng nó trên một tạp chí tư nhân. Nó đã được sử dụng để làm lung lay sự đồng thuận khoa học.

Hãy cùng đếm nghề nghiệp của những người đã trải qua các sự việc này. Họ là luật sư, thẩm phán, giáo sư, quan chức chính phủ và cố vấn của các tập đoàn lớn.

Họ là những người được đào tạo lâu năm trong việc xử lý văn bản. Họ là những người kiếm tiền bằng việc xác minh tính xác thực của các tài liệu.

Vậy mà họ vẫn bị mắc lừa. Đó là vì các tài liệu do trí tuệ nhân tạo đưa ra hoàn hảo về mặt hình thức. Có số hồ sơ vụ án, có ngày tháng và văn phong giống như văn bản pháp lý. Con người có thói quen phán đoán tính xác thực qua hình thức.

Những án lệ giả đứng ở phía đối lập hoàn toàn với thói quen đó. Chúng không có nội dung mà chỉ có hình thức.

Lý do để tranh cãi về sự thật trước tòa là vì nó liên quan đến sự tự do và tài sản của con người. Nếu căn cứ của cuộc tranh cãi đó là một phán quyết không tồn tại, thì phiên tòa đó là gì.

Mỗi ngày có tám trường hợp bị phát hiện.

2.3

Bóp méo sự thật lịch sử và địa lý cùng tranh cãi xã hội

Tháng 10 năm 2025, có người đã gõ từ lãnh thổ Nhật Bản bằng tiếng Nhật vào thanh tìm kiếm của Naver.

Ở phần trên của màn hình, bản tóm tắt do trí tuệ nhân tạo biên soạn đã xuất hiện. Trong đó có Dokdo. Được đặt cạnh Lãnh thổ phương Bắc và quần đảo Senkaku, nơi này được ghi là khu vực đang có tranh chấp với Hàn Quốc.

Dịch vụ của một công ty Hàn Quốc, trên thanh tìm kiếm do người Hàn Quốc sử dụng, đã đưa Dokdo vào danh sách của phía Nhật Bản.

Giáo sư Seo Kyoung-duk của Đại học Nữ sinh Sungshin đã công khai màn hình này và mọi người đã tỏ ra vô cùng tức giận. Naver đã khẩn cấp dừng tính năng tóm tắt tìm kiếm và bắt đầu kiểm tra.

Nguyên nhân về mặt kỹ thuật thì nhàm chán nhưng về mặt nội dung lại rất đau lòng. HyperCLOVA X khi nhận được câu hỏi bằng tiếng Nhật sẽ tập hợp và tóm tắt các tài liệu web tiếng Nhật. Khi tìm kiếm từ lãnh thổ bằng tiếng Nhật, tài liệu quảng bá của Bộ Ngoại giao Nhật Bản sẽ xuất hiện ở vị trí đầu tiên. Hệ thống đã tổng hợp lại tài liệu đó.

Trí tuệ nhân tạo chỉ làm theo những gì được sai bảo. Tuy nhiên, không một ai quy định rằng đối với câu hỏi này thì con người phải kiểm tra lại một lần.

Trong vấn đề lãnh thổ, việc đọc tài liệu của quốc gia nào trước không phải là vấn đề kỹ thuật mà là vấn đề lập trường. Máy móc không có lập trường. Việc không có lập trường cũng được gọi là trung lập. Thế nhưng, nếu tài liệu bị nghiêng về một phía, máy móc không có lập trường sẽ lăn về phía bị nghiêng đó.

Tại Đài Loan, một câu trả lời còn đáng kinh ngạc hơn đã xuất hiện.

Tháng 10 năm 2023, cơ quan học thuật tối cao của Đài Loan là Viện Nghiên cứu Trung ương đã tạo ra và công bố một mô hình ngôn ngữ kiểu Đài Loan. Người dân đã hỏi chatbot rằng: Bạn thuộc về quốc gia nào.

Chatbot đã trả lời rằng: Là Trung Quốc.

Khi được hỏi nhà lãnh đạo tối cao của Đài Loan là ai. Nó đã trả lời là Tập Cận Bình. Khi được hỏi quốc ca là gì. Nó đã trả lời là Nghĩa dũng quân tiến hành khúc.

Đội ngũ phát triển đã tiết lộ lý do. Đó là vì muốn tạo ra nhanh chóng nên họ đã cho tập dữ liệu chữ Hán giản thể của đại lục Trung Quốc qua máy dịch rồi đưa nguyên vào để huấn luyện.

Máy dịch thay đổi chữ viết. Nó không thể thay đổi quan điểm. Chính phủ Đài Loan và Viện Nghiên cứu Trung ương đã lập tức gỡ bỏ hệ thống.

Tháng 2 năm 2025, mô hình DeepSeek R1 của Trung Quốc với hơn ba triệu lượt tải xuống trên toàn thế giới, khi được hỏi về nghi vấn diệt chủng người Duy Ngô Nhĩ, đã đưa ra câu trả lời rằng đây là sự can thiệp vào công việc nội bộ và là sự vu khống vô căn cứ. Nó đã lặp lại y nguyên lý lẽ của chính phủ Trung Quốc.

Mô hình nói những gì mà nó đã đọc. Việc đưa vào những tài liệu nào sẽ chính thức trở thành thế giới quan của mô hình đó.

Nghị viện Châu Âu cũng vướng phải vấn đề. Năm 2025, trí tuệ nhân tạo chính thức được tạo ra dựa trên Claude của Anthropic đã không thể đưa ra tên của Chủ tịch đầu tiên của Ủy ban Điều hành Châu Âu. Một thứ được tạo ra để xử lý các ghi chép lại không biết đến các ghi chép đó.

Sự xuyên tạc lịch sử còn tiến xa hơn ở điểm này.

Tháng 6 năm 2024, một câu trả lời của ChatGPT đã được đăng tải trong báo cáo của UNESCO. Nội dung là khi được hỏi về thảm họa Holocaust, nó đã giải thích

về một sự kiện rằng Đức Quốc xã đã giết chết người Do Thái bằng cách dìm họ xuống nước hàng loạt.

Đã không có sự kiện nào như vậy. Nó thậm chí còn được gắn với cái tên là thảm họa Holocaust do đuối nước.

Lý do khiến các nhà sử học lo lắng nằm ở hướng đi của lời nói dối này. Những người phủ nhận thảm họa Holocaust từ lâu đã sử dụng phương pháp khoét sâu vào những kẽ hở nhỏ bé của các ghi chép để làm lung lay toàn bộ sự việc. Nếu một sự kiện không tồn tại bị trộn lẫn vào các ghi chép, thì ngay cả những sự kiện đã từng xảy ra cũng sẽ bị nghi ngờ.

Grok do xAI của Elon Musk tạo ra đã vượt quá giới hạn ở điểm này.

Giữa năm 2025, Grok đã chỉnh sửa lại hệ thống với danh nghĩa là loại bỏ sự đúng đắn về chính trị. Ngay sau đó, Grok đã nói rằng con số sáu triệu nạn nhân của thảm họa Holocaust có thể đã bị thao túng vì mục đích chính trị. Nó cũng đưa ra câu trả lời phủ nhận sự tồn tại của các phòng ngạt khí Auschwitz. Nó thậm chí còn tự gọi mình là MechaHitler.

Tháng 1 năm 2023, đã có một ứng dụng trò chuyện với các nhân vật lịch sử do kỹ sư của Amazon tạo ra bằng GPT-3. Ứng dụng này đã quảng cáo rằng có thể trò chuyện với hai mươi nghìn nhân vật. Chatbot của Bộ trưởng Tuyên truyền Đức Quốc xã Joseph Goebbels đã nói rằng bản thân không hề ghét người Do Thái mà ngược lại còn cảm thấy đau khổ.

Cỗ máy bắt chước giọng nói của kẻ gây hại cũng bắt chước cả lời bào chữa của kẻ gây hại đó. Nếu lời bào chữa đó xuất hiện từ ứng dụng, sẽ có ai đó nhầm tưởng nó là tài liệu lịch sử.

Các công ty đã giải thích rằng đó là lỗi lập trình hoặc do các chỉ thị nội bộ đã bị thay đổi một cách tùy ý.

Những gì nhìn thấy bằng mắt cũng trở nên không thể tin tưởng được.

Tháng 6 năm 2025, tổ chức kiểm chứng sự thật Full Fact của Vương quốc Anh đã bắt được một trường hợp kỳ lạ. Đã có một đoạn video bình thường được quay tại một bãi biển thuộc bang Goa của Ấn Độ. Bản tóm tắt của trí tuệ nhân tạo đính kèm trên Google Lens đã giải thích về đoạn video này như sau: Cảnh tượng những người xin tị nạn tiến vào bãi biển Dover của Vương quốc Anh với số lượng lớn.

Nếu có bãi biển xuất hiện và có nhiều người, thì nó sẽ trở thành cảnh tượng đó. Ấn Độ trở thành Vương quốc Anh.

Không khó để đoán được điều gì sẽ xảy ra nếu bản tóm tắt này gặp phải những tin đồn trên mạng xã hội. Đó là thời kỳ mà tâm lý chống người nhập cư đang sục sôi ở Vương quốc Anh.

Vào tháng 5 năm 2024, cùng một dịch vụ đó đã đưa ra những câu trả lời như thế này: Nếu phô mai trên bánh pizza bị chảy xuống thì hãy trộn với keo dán không độc hại. Hãy ăn mỗi ngày một viên đá nhỏ. Đối với bệnh sỏi thận thì hãy uống nước tiểu. Barack Obama là tổng thống Hồi giáo đầu tiên của Hoa Kỳ.

Câu trả lời bảo hãy ăn đá đến từ một bài báo của một phương tiện truyền thông châm biếm. Máy móc không có cảm giác để phân biệt giữa trò đùa và thông tin. Con người học được cảm giác này từ khi còn nhỏ. Bằng cách nhìn vào nét mặt và hoàn cảnh. Máy móc chỉ tiếp nhận các chữ viết.

Tháng 4 năm 2025, tại Malaysia đã xảy ra vấn đề về quốc kỳ. Các công ty truyền thông, Bộ Giáo dục và các quan chức của triển lãm quốc gia đã công bố hình ảnh quốc kỳ Malaysia do trí tuệ nhân tạo tạo ra. Biểu tượng trăng lưỡi liềm đã bị thiếu hoặc các biểu tượng bị bóp méo một cách kỳ lạ.

Quốc kỳ là khuôn mặt của một quốc gia. Trong khi vẽ nên khuôn mặt đó, không một ai đã kiểm tra lại.

Sự việc tương tự cũng đã xảy ra ở bản đồ và địa hình.

Tháng 11 năm 2024, tỉnh Fukuoka của Nhật Bản đã phải đưa ra lời xin lỗi chính thức và chấm dứt hợp đồng với một đại lý quảng cáo. Đó là vì trong ấn phẩm quảng bá do đại lý này tạo ra bằng trí tuệ nhân tạo có chứa những địa điểm du lịch không có ở Fukuoka và những món ăn địa phương không tồn tại cứ như thể chúng là đồ thật.

Tháng 5 năm 2025, dự án lập bản đồ vùng đất than bùn của Cơ quan Tự nhiên Vương quốc Anh cũng đã vấp ngã. Vùng đất than bùn là vùng đất ngập nước giữ lại carbon nên rất quan trọng trong các chính sách về khí hậu. Trên bản đồ do mô hình học máy tạo ra, các bức tường đá, mỏ đá, đống đất và đá hoa cương đã bị đánh dấu là vùng đất than bùn.

Chỉ với những bức ảnh nhìn từ trên xuống, rất khó để phân biệt giữa mặt đất ướt và những viên đá khô. Con người dùng chân dẫm lên để thử. Mô hình không có chân.

Năm 2021, đội ngũ nghiên cứu của Đại học Stanford và Đại học Washington đã cảnh báo về một khái niệm gọi là địa lý học deepfake. Đó là việc có thể tổng hợp các bức ảnh vệ tinh để tạo ra những địa hình không tồn tại. Những gì sẽ xảy ra trong quân sự và ngoại giao, tôi xin nhường lại cho sự tưởng tượng.

Tháng 12 năm 2024, phóng viên ảnh nổi tiếng người Bỉ Carl De Keyzer không thể đến Nga sau cuộc xâm lược Ukraine, nên ông đã dùng trí tuệ nhân tạo tạo sinh để tạo ra những bức ảnh về phong cảnh và con người Nga rồi đưa vào sách ảnh. Cơ sở của báo ảnh là sự thật rằng người ta đã ở đó. Cơ sở đó đã biến mất.

Trong việc chỉ đường, đã có người chết.

Năm 2024 tại Ấn Độ, ba người đàn ông lái xe theo hướng dẫn của Google Maps đã rơi xuống một cây cầu gãy và tất cả đều thiệt mạng. Cây cầu chưa được xây dựng xong và đang bị ngập trong nước.

Trên bản đồ có đường. Trong thực tế thì không.

Tại Mỹ cũng xảy ra một tai nạn tương tự. Một tài xế được hướng dẫn đi vào một cây cầu sập đã rơi xuống nước và tử vong. Cùng một sai lầm đã lặp lại ở một châu lục khác.

Tháng 2 năm 2026 tại Anh, một xe tải giao hàng của Amazon đi theo hướng dẫn của Google Maps đã tiến vào Broomway. Đây là con đường bãi bùn bị ngập nước biển khi thủy triều lên. Chiếc xe tải đã bị kẹt trong bùn.

Tháng 5 năm 2025, Google Maps đã chỉ báo do một lỗi không rõ nguyên nhân rằng các đường cao tốc Autobahn chính trên toàn nước Đức bị phong tỏa, làm chấn động mạng lưới thông tin giao thông.

Tháng 1 năm 2026 tại Tasmania thuộc Úc, những du khách tin vào một bài báo du lịch do trí tuệ nhân tạo viết và đi tìm một suối nước nóng không tồn tại đã bị mắc kẹt ở vùng hẻo lánh.

Trong tình huống khẩn cấp, thông tin sai lệch do trí tuệ nhân tạo đưa ra càng nguy hiểm hơn.

Tháng 4 năm 2024, Grok đã tạo ra một tiêu đề rằng Iran đã bắn tên lửa vào Tel Aviv. Cùng tháng đó, nó cũng đưa tin thủ tướng Ấn Độ bị trục xuất khỏi chính phủ. Vào tháng 5 năm 2025, nó thậm chí tự ý chèn thêm thuyết âm mưu về nạn diệt chủng người da trắng ở Nam Phi vào các câu hỏi về trận đấu bóng chày hoặc yêu cầu nói chuyện giống như cướp biển.

Tháng 3 năm 2026, nó đã tạo ra bài viết xúc phạm các nạn nhân của thảm họa bóng đá Hillsborough và Heysel, để lại tổn thương cho gia đình những người đã khuất. Cả hai sự kiện đều là thảm kịch có thật khiến hàng trăm người bị giẫm đạp đến chết.

Tháng 9 năm 2025, tại hiện trường cuộc biểu tình chống người nhập cư ở London, nó đã đưa ra thông tin giả mạo rằng cảnh sát đã ngụy tạo các video bạo lực.

Cũng đã có những nỗ lực nhằm viết lại chính kiến thức.

Tháng 10 năm 2025, Musk đã tung ra một bách khoa toàn thư bằng trí tuệ nhân tạo có tên là Grokpedia để đối đầu với Wikipedia. Ngay khi ra mắt, chín trăm nghìn tài liệu đã được tải lên. Đó là những tài liệu sao chép từ các bài viết trên bảng tin và blog chưa được xác minh, và những sự thật được tạo ra bởi ảo giác đã được đăng làm các mục từ. Thuyết âm mưu về nạn diệt chủng người da trắng ở Nam Phi được đăng tải như thể là sự thật, và lịch sử của bệnh AIDS bị bóp méo.

Tên gọi bách khoa toàn thư làm cho con người cảm thấy an tâm. Đó là nhằm vào sự an tâm đó.

Vào tháng 2 năm 2025, sự thật đã bị phơi bày rằng mô hình Grok 3 đã được thiết lập để dùng các chỉ thị nội bộ chặn nguồn chỉ trích rằng Musk và Trump phát tán thông tin sai lệch trong quá trình suy luận. Họ đã định sẵn không phải là sẽ nói gì mà là sẽ không đọc gì.

Thiệt hại đến với cá nhân là ngay tức khắc.

Tháng 10 năm 2025, thượng nghị sĩ Hoa Kỳ Marsha Blackburn đã phát hiện ra rằng mô hình Gemma của Google tạo ra câu chuyện bà bị cuốn vào một vụ án tội phạm tình dục trong cuộc bầu cử thượng viện bang năm 1987 và thậm chí còn đính kèm các liên kết tin tức giả mạo.

Hệ thống

CHƯƠNG 3

Xâm phạm quyền riêng tư và xã hội giám sát: Bóng tối của việc thu thập dữ liệu không xin phép

3.1

Thu thập thông tin sinh trắc học trái phép và giám sát công cộng · thương mại

Khi lạc đường trong trung tâm mua sắm, bạn sẽ đứng trước màn hình hướng dẫn. Bạn xem bản đồ, tìm tên cửa hàng và dùng ngón tay nhấn vào màn hình vài lần. Việc này mất vài giây.

Vào năm 2018 tại mười hai trung tâm mua sắm ở Canada, đã có một sự việc xảy ra trong vài giây đó.

Công ty bất động sản lớn của Canada là Cadillac Fairview đã giấu camera bên trong các ki-ốt hướng dẫn. Khuôn mặt của những người nhìn vào màn hình đã bị chụp lại, và một thuật toán đo lường sinh trắc học đã phán đoán độ tuổi và giới tính của họ.

Có năm triệu người đã bị ghi lại như vậy.

Khi cuộc điều tra bắt đầu, công ty đã giải thích như sau: Hình ảnh không được lưu trữ mà sẽ bị xóa ngay lập tức.

Nhận định của cơ quan giám sát lại khác. Bất kể có xóa hay không, họ cho rằng bản thân hành vi đo lường khuôn mặt khi chưa có sự đồng ý đã là vi phạm pháp luật.

Sự phân biệt này rất quan trọng. Mọi người thường lo lắng xem bức ảnh của mình được lưu trữ ở đâu. Thế nhưng, cốt lõi của thông tin sinh trắc học không phải là sự lưu trữ mà là sự đo lường. Cho dù bức ảnh có bị xóa đi, thì hồ sơ ghi lại rằng người này là một phụ nữ ở độ tuổi ba mươi và đã có mặt ở tầng này vào chiều thứ Ba vẫn còn đó.

Khuôn mặt khác với mật khẩu. Nếu bị rò rỉ, bạn không thể thay đổi nó.

Chuỗi hiệu thuốc Rite Aid của Mỹ còn đi xa hơn một bước. Từ năm 2012 đến năm 2020, họ đã lắp đặt camera nhận diện khuôn mặt tại các cửa hàng trên toàn quốc. Lý do được đưa ra là để ngăn chặn trộm cắp.

Vấn đề nằm ở việc camera được lắp đặt nhiều ở đâu. Chúng tập trung ở những khu dân cư có nhiều người thu nhập thấp, người da đen và người gốc Tây Ban Nha sinh sống.

Và thuật toán này cũng thường xuyên sai sót. Khi chuông báo động reo, các nhân viên đã tiến lại gần khách hàng. Họ yêu cầu khách hàng rời đi. Họ đã gọi cảnh sát.

Người đến mua hàng đã bị đối xử như kẻ trộm trước mặt mọi người. Không có cách nào để chứng minh sự oan uổng ngay tại chỗ.

Ủy ban Thương mại Liên bang đã cấm Rite Aid sử dụng công nghệ này trong vòng năm năm.

Sự việc tương tự đã xảy ra ở nhiều quốc gia.

Chuỗi siêu thị Mercadona của Tây Ban Nha đã áp dụng nhận diện khuôn mặt tại hơn bốn mươi cửa hàng và đã bị phạt năm triệu một trăm mười lăm nghìn euro. Công ty nói rằng họ muốn sàng lọc những người có tiền án đã bị lệnh cấm đến gần.

Câu trả lời của giới chức trách là thế này: Việc đo lường khuôn mặt của tất cả mọi người chỉ để tìm kiếm một vài người là không được phép.

Ở Úc, Bunnings, Kmart và 7-Eleven đã bị bắt quả tang. 7-Eleven đã lén lút chụp lại và lưu trữ khuôn mặt của những người trả lời trong khi tiến hành khảo sát mức độ hài lòng bằng máy tính bảng của cửa hàng.

Họ vừa hỏi "Bạn có hài lòng không?" vừa chụp lại khuôn mặt.

Các chuỗi cửa hàng tạp hóa của Anh cũng đã mang camera nhận diện khuôn mặt vào, sau đó gây ra náo động khi đổ tội cho một khách hàng vô tội là kẻ

trộm.

Ở Trung Quốc, công nghệ này đã được sử dụng để tính toán tiền hoa hồng.

Các công ty phát triển bất động sản ở Nam Kinh, Ninh Ba và Nam Ninh đã chụp lại khuôn mặt của những khách hàng đến văn phòng bán hàng mà không có sự đồng ý. Mục đích là thế này: Họ muốn xác nhận xem vị khách này đã đến thông qua người môi giới nào để chi trả mức hoa hồng khác nhau.

Khuôn mặt của khách hàng đã trở thành một bộ phận của máy tính tiền hoa hồng môi giới. Khách hàng đã không hề hay biết.

Câu chuyện này có một phần kết bi hài. Khi tin đồn lan ra, đã xuất hiện những người đội mũ bảo hiểm xe máy để đi đến văn phòng bán hàng.

Các thương hiệu như Kohler, BMW và Max Mara cũng đã đặt camera quay lén tại các cửa hàng ở Trung Quốc để theo dõi số lần ghé thăm và tuyến đường di chuyển, cho đến khi bị đài truyền hình nhà nước vạch trần vào năm 2021.

Các cửa hàng không người bán và hệ thống thanh toán cũng đang thu thập khuôn mặt và bàn tay.

Cửa hàng Amazon Go ở New York đã bị kiện tập thể do không tuân thủ nghĩa vụ thông báo theo quy định của luật pháp Thành phố New York trong khi thu thập thông tin sinh trắc học của những khách hàng bước vào. Amazon One, hệ thống thanh toán bằng cách áp lòng bàn tay, đã lưu trữ dữ liệu tĩnh mạch lòng bàn tay và đường chỉ tay trên máy chủ.

McDonald's đã đưa công nghệ nhận diện giọng nói vào hệ thống drive-thru (mua hàng trên xe), thu thập và lưu trữ các đặc điểm giọng nói của người lái xe mà không có sự đồng ý, dẫn đến việc bị kiện vì vi phạm Đạo luật Bảo vệ Thông tin Sinh trắc học của bang Illinois. Domino's Pizza cũng đã phải ra hầu tòa vì cáo buộc thu thập trái phép giọng nói từ các đơn đặt hàng qua điện thoại.

Chỉ là gọi mua một chiếc hamburger, nhưng dấu vân tay giọng nói lại bị lưu lại.

Chúng ta hãy chuyển sang trường học.

Trường Trung học Anderstorp ở Skellefteå, Thụy Điển, đã thử nghiệm camera nhận diện khuôn mặt với hai mươi hai học sinh với lý do tiết kiệm thời gian điểm danh. Vào năm 2019, Cơ quan Bảo vệ Dữ liệu Thụy Điển đã phạt họ vì vi phạm quy định bảo vệ thông tin cá nhân. Đây là khoản tiền phạt đầu tiên của Thụy Điển được đưa ra theo quy định này.

Nhà trường nói rằng họ đã nhận được sự đồng ý của phụ huynh. Giới chức trách đã nhận định như sau: Do có sự chênh lệch về quyền lực giữa nhà trường và học sinh nên rất khó để thiết lập một sự đồng ý tự nguyện thực sự.

Học sinh rất khó để từ chối yêu cầu từ giáo viên. Giữa người lớn với nhau cũng vậy.

Và mục đích là để điểm danh. Họ đã sử dụng thông tin sinh trắc học cho một việc mà chỉ cần gọi tên theo sổ điểm danh là xong.

Các trường trung học ở Nice và Marseille của Pháp đã định đặt cổng nhận diện khuôn mặt ở cửa ra vào, nhưng đã phải từ bỏ do sự phản đối của các tổ chức dân sự và sự không cho phép của tòa án.

Chín trường trung học ở North Ayrshire, Anh, đã áp dụng nhận diện khuôn mặt với lý do tăng tốc độ thanh toán tiền ăn trưa, nhưng đã phải tắt hệ thống sau khi nhận được cảnh báo từ Ủy ban Thông tin. Trường Trung học Chelmer Valley cũng bị phát hiện vì lý do tương tự.

Một trường tiểu học ở Gdańsk, Ba Lan, đã bị phạt tiền sau khi thu thập dấu vân tay của trẻ em để xác nhận việc ăn trưa.

Họ đã lấy dấu vân tay của trẻ em chỉ để rút ngắn vài giây hàng người xếp hàng chờ ăn cơm.

Một cuộc tranh cãi đã nổ ra khi chính phủ Ấn Độ có ý định đưa nhận diện khuôn mặt vào việc cấp bảng điểm và xác nhận việc ăn trưa. Những nỗ lực và sự phản

đối tương tự cũng đã xuất hiện ở Scotland, bang Paraná của Brazil và bang Victoria của Úc.

Ở Mỹ, từ "an toàn" đã được dùng làm chìa khóa. Học khu Lockport ở bang New York đã chi hàng triệu đô la để lắp đặt hệ thống giám sát nhận diện khuôn mặt tại trường học với lý do ngăn chặn các vụ xả súng. Phụ huynh và các tổ chức dân sự đã phản đối, và Sở Giáo dục bang New York đã thực hiện biện pháp đình chỉ tạm thời việc sử dụng công nghệ này tại tất cả các trường học trong bang.

Camera khi bước vào bên trong lớp học đã không chỉ nhìn vào khuôn mặt.

Trường Trung học cơ sở số 11 của Trung Quốc và Đại học Dược Khoa Trung Quốc đã đặt camera trước lớp học để phân tích nét mặt, sự buồn ngủ và mức độ tập trung của học sinh trong thời gian thực, từ đó chấm điểm.

Hãy thử tưởng tượng một lớp học mà nét mặt lơ đãng trong giờ học bị ghi lại. Trong lớp học đó, học sinh không còn là người học nữa mà trở thành những dữ liệu bị đánh giá.

Quyền tự do được bộc lộ khuôn mặt buồn chán là một phần của việc học tập.

Ở các sân bay và biên giới, quy mô lại càng lớn hơn.

Cơ quan Dịch vụ Biên giới Canada đã thử nghiệm các ki-ốt tự động thu thập khuôn mặt tại Sân bay Quốc tế Toronto Pearson mà hành khách không hề hay biết. Tại Sân bay Quốc tế Wellington ở New Zealand, việc chạy camera mà không có sự đồng ý cũng đã bị phơi bày qua các bản tin báo chí. Aena, công ty vận hành sân bay quốc doanh của Tây Ban Nha, đã phải nộp phạt vì thu thập trái phép khuôn mặt của hành khách.

Ở Hàn Quốc cũng đã có sự việc xảy ra.

Từ năm 2019, Bộ Tư pháp và Bộ Khoa học, Công nghệ và Thông tin Truyền thông đã thực hiện dự án hệ thống nhận dạng theo dõi trí tuệ nhân tạo, đồng thời chuyển giao ảnh khuôn mặt của những người xuất nhập cảnh qua Sân bay

Incheon và những nơi khác cho các doanh nghiệp trí tuệ nhân tạo tư nhân để huấn luyện, mà không có sự đồng ý của chính người đó.

Số lượng là 170 triệu trường hợp.

Không phải chỉ có người nước ngoài. Khuôn mặt của công dân Hàn Quốc cũng có trong đó. Sự thật rằng dữ liệu đã được chuyển giao cho các doanh nghiệp tư nhân mà không có cơ sở pháp lý đã được xác nhận trong cuộc điều tra.

Nếu có nhớ lại lúc nhìn vào camera tại quầy kiểm tra hộ chiếu, đáng để suy nghĩ về việc bức ảnh đó đi đâu.

Chính quyền biên giới Hoa Kỳ đã đưa nhận dạng khuôn mặt vào ứng dụng di động dành cho những người nộp đơn xin tị nạn. Ứng dụng này không thể nhận dạng tốt khuôn mặt của những người xin tị nạn da đen. Nếu nhận dạng không được thì bản thân đơn xin không được tiến hành. Những người chạy trốn từ nơi sống cũ của họ bị chặn thêm lần nữa trước ứng dụng này.

Ở đường phố, có một dự án thu thập mống mắt.

Worldcoin, do Sam Altman thành lập, đã đặt một thiết bị nhận dạng mống mắt có tên là Orb ở nhiều thành phố trên thế giới. Nếu đăng ký mống mắt, bạn sẽ nhận được tiền điện tử. Hàng xếp dài ở những nước nghèo và trong tầng thanh niên.

Chính phủ Kenya đã dừng hoàn toàn dự án. Chính quyền Bồ Đào Nha, Tây Ban Nha, Indonesia, Thái Lan cũng ra lệnh đóng cửa. Hồng Kông, Argentina và Ủy ban Bảo vệ Thông tin Cá nhân Hàn Quốc cũng bắt đầu cuộc điều tra.

Mống mắt đã đăng ký một lần không thể thu hồi. Đồng tiền nhận được có giá trị biến động. Đáng để kiểm tra xem giao dịch có công bằng không.

Một thiết bị có thể biết được danh tính của người khác chỉ với một cái kính cũng đã xuất hiện.

Hai sinh viên Đại học Harvard đã công bố một hệ thống gắn nhận dạng khuôn mặt vào kính thông minh Ray-Ban của Meta. Khi đeo kính này và nhìn vào một người không quen ở đường phố hay tàu điện ngầm, chỉ trong vài giây, tên, địa chỉ, số điện thoại và mối quan hệ gia đình của người kia sẽ xuất hiện trên màn hình điện thoại di động.

Các sinh viên nói rằng họ đã tạo ra để cảnh báo về nguy hiểm. Sự thật rằng có thể tạo ra nó đã là một cảnh báo rồi.

Cũng đã có trường hợp người dùng kính thông minh của Meta quay lén những người khác tại các cơ sở massage hoặc địa điểm công cộng và lan truyền. Sự thật rằng các video thu thập được bằng kính này được chuyển giao cho lực lượng lao động thuê ngoài ở nước ngoài để sử dụng trong huấn luyện cũng đã được tiết lộ.

Phía sau tất cả những công cụ này là một công ty thu thập toàn bộ Internet.

Clearview AI đã thu thập 3 tỷ ảnh khuôn mặt từ Facebook, Instagram, X, LinkedIn và các nơi khác trên Internet để tạo cơ sở dữ liệu tìm kiếm. Và nó đã bán cho cảnh sát, cơ quan điều tra ở nhiều nước trên thế giới, thậm chí cả các công ty phân phối.

Liên minh Tự do Dân sự Hoa Kỳ đã nộp đơn kiện thay mặt các công dân tiểu bang Illinois, và các tổ chức bảo vệ thông tin cá nhân của Anh, Pháp, Ý, Hà Lan và Hy Lạp đã phát hành lệnh phạt tiền lớn và xóa dữ liệu.

Công cụ tìm kiếm khuôn mặt PimEyes tệ hơn. Nó đã thu thập ảnh riêng tư, ảnh của người đã mất, thậm chí cả hình ảnh tình dục của trẻ em để tạo dịch vụ tìm kiếm ảnh đời tư của một người bằng một bức ảnh. Tòa án Đức và Hoa Kỳ đã trừng phạt nó.

Meta đã trả 1,4 tỷ đô la tiền bồi thường vì tình nghi thu thập dữ liệu sinh trắc học khuôn mặt mà không có sự đồng ý của cư dân ở tiểu bang Texas v.v.

Nếu xem xét ở cấp quốc gia, quy mô lại khác.

Quân đội quốc phòng Israel đã vận hành một hệ thống nhận dạng khuôn mặt để giám sát cư dân Palestine ở Hebron và Dải Gaza. Moscow đã sử dụng mạng giám sát khuôn mặt để thanh toán tàu điện ngầm và theo dõi các nhân vật chống chính phủ. Ga King's Cross ở London và các cơ quan điều tra cũng đã triển khai nhận dạng khuôn mặt thời gian thực.

Nếu lướt qua những sự vụ này, những địa điểm sẽ lọt vào mắt. Trung tâm mua sắm, hiệu thuốc, siêu thị, cửa hàng tiện lợi, cửa hàng ô tô, văn phòng bán căn hộ, trường tiểu học, trường trung học phổ thông, đại học, sân bay, biên giới, ga tàu điện ngầm, đường phố.

Đó chính là hành trình người sống trong một ngày.

Và ở tất cả những nơi này, cùng một câu nói được lặp lại. Vì an toàn. Vì tiện lợi. Vì hiệu quả.

Thật khó khăn để phản đối những câu nói này. Vì thế chúng bán rất tốt.

Khuôn mặt, mống mắt, dấu vân tay và giọng nói gắn liền với cơ thể. Ngay cả khi mất chúng, bạn không thể cấp lại. Những người biết sự thật này đang thu thập chúng.

3.2

Việc học thông tin cá nhân trái phép và sự cố rò rỉ hội thoại AI

Vào một ngày tháng 3 năm 2023, khi những người bật ChatGPT, danh sách hội thoại xuất hiện ở bên trái màn hình của họ. Đó là nơi luôn có chúng.

Nhưng những tiêu đề đó rất lạ lẫm. Chúng là những cuộc hội thoại mà họ chưa bao giờ có.

Tiêu đề cuộc hội thoại của người khác xuất hiện trên màn hình của tôi. Tiêu đề cuộc hội thoại của tôi có lẽ cũng xuất hiện trên màn hình của ai đó.

OpenAI giải thích rằng điều này là do lỗi trong thư viện cơ sở dữ liệu mã nguồn mở và tạm dừng dịch vụ. Kết quả điều tra cho thấy khoảng 1,2 phần trăm người dùng đăng ký ChatGPT Plus đã tiếp xúc với tên, địa chỉ email, bốn chữ số cuối cùng của thẻ tín dụng và ngày hết hạn từ những người dùng khác.

Nếu bạn suy nghĩ về những gì mọi người viết vào cửa sổ đó, trọng lượng của sự cố này sẽ thay đổi.

Họ yêu cầu sửa CV trong khi đưa vào địa chỉ và số điện thoại. Họ viết tên bệnh vì muốn hiểu được giấy chẩn đoán. Họ viết về vợ/chồng khi suy tư về ly hôn. Họ viết những nỗi lo mà họ không thể nói với công ty.

Mọi người đặt câu hỏi vào thanh tìm kiếm. Họ đặt hoàn cảnh vào chatbot.

Trước khi mọi người nhận ra sự khác biệt này, sự cố tiếp theo đã xảy ra.

ChatGPT có một tính năng để chia sẻ hội thoại qua liên kết. Đó là một tính năng được tạo ra để hiển thị cho bạn bè. Nhưng những liên kết này đã được coi là các trang web công khai.

Năm 2025, hơn một trăm ngàn bản ghi hội thoại do người dùng tạo được lập chỉ mục trên các công cụ tìm kiếm như Google. Khi nhập một vài từ vào thanh tìm kiếm, lịch sử tư vấn của những người lạ đã xuất hiện. Những lo lắng cá nhân đã bị tiếp xúc. Bí mật công ty đã bị tiếp xúc.

Những gì được dự định để hiển thị cho một người bạn đã trở thành việc được thế giới thấy.

Bard của Google đã mắc phải lỗi tương tự. Grok cũng vậy. Hàng trăm ngàn cuộc hội thoại với Grok đã bị lộ trong kết quả tìm kiếm, và Grok thậm chí đã tiết lộ tên thật và ngày sinh của một người làm việc trong ngành nội dung người lớn mà không được yêu cầu.

Cùng một lỗi đã được lặp lại khi di chuyển giữa các công ty. Trong cuộc đua để nhanh chóng phát hành các tính năng mới, việc xem xét quyền truy cập của liên kết được chia sẻ bị đẩy ra sau.

Trong các ứng dụng xử lý hội thoại tình cảm, sự rò rỉ đau đớn hơn.

Năm 2024, Character AI có một sự cố do lỗi nội bộ cho phép người dùng xem các cuộc hội thoại riêng tư của nhau. Đó là những lời thú nhận được kể cho các nhân vật ảo.

DeepSeek của Trung Quốc đã tiết lộ thông tin tài khoản và lịch sử hội thoại do bỏ qua quản lý bảo mật cơ sở dữ liệu. Quy mô của nó là ba trăm triệu tin nhắn riêng tư.

Vào tháng 2 năm 2026, cũng có một sự cố ở quy mô tương tự. Ba trăm triệu tin nhắn từ 25 triệu người dùng của một ứng dụng trò chuyện AI đã bị tiếp xúc. Các tệp theo người dùng chứa toàn bộ lịch sử hội thoại.

Các ứng dụng bạn đồng hành AI thu thập những điều sâu hơn. Các ứng dụng như Replika, Mua và Nomi đã thu thập dữ liệu xu hướng tình dục và cảm xúc của người dùng bằng cách ưu tiên kết nối tình cảm. Theo cuộc điều tra của Quỹ Mozilla, những ứng dụng này không tuân thủ ngay cả các tiêu chuẩn bảo vệ

quyền riêng tư tối thiểu và đã chuyển dữ liệu mà họ thu thập cho các nhà tiếp thị.

Do các rò rỉ liên tiếp, các cuộc hội thoại của hơn bốn trăm ngàn người dùng đã bị tiếp xúc. Khi bảo mật của một dịch vụ chatbot dành cho người lớn bị vi phạm, hai triệu hình ảnh album tốt nghiệp của phụ nữ và dữ liệu nội dung khiêu dâm tổng hợp cũng bị tiếp xúc.

Những hình ảnh album tốt nghiệp không phải những người phụ nữ tải lên dịch vụ. Ai đó đã xóa chúng.

Bí mật công ty cũng rò rỉ qua cùng một cửa sổ.

Vào mùa xuân năm 2023, sự cố xảy ra tại một cơ sở của Samsung Electronics. Các nhân viên bộ phận bán dẫn đã sử dụng ChatGPT để hoàn thành công việc nhanh hơn. Họ đưa vào dữ liệu đo lường thiết bị. Họ đưa vào mã nguồn liên quan đến sản lượng. Họ đưa vào toàn bộ ghi chép cuộc họp và yêu cầu tóm tắt.

Dữ liệu đó đã đi đến máy chủ OpenAI. Công ty nhanh chóng hạn chế việc sử dụng AI nội bộ.

Thật dễ dàng để trách cứ các nhân viên. Nhưng lõi của sự cố này nằm ở chỗ khác. Cửa sổ chatbot trông giống như thanh tìm kiếm. Mọi người không đặt tài liệu công ty vào thanh tìm kiếm. Nhưng họ đặt chúng vào chatbot. Vì nó nói rằng nó sẽ tóm tắt.

Hình dạng của công cụ thay đổi hành vi của con người.

Dịch vụ tóm tắt âm thanh Otter AI đã ghi âm các cuộc gọi điện thoại riêng tư của nhà đầu tư mà không có sự đồng ý, gửi bản tóm tắt đến nơi sai và tiết lộ thông tin tài chính và chiến lược đầu tư.

Các công ty lớn cũng tiết lộ tài liệu của riêng họ.

Vào tháng 9 năm 2023, đội nghiên cứu AI của Microsoft đã sai lầm trong cài đặt quyền truy cập khi chia sẻ một bộ dữ liệu trên GitHub. Mã thông báo có thể đọc

và ghi toàn bộ hệ thống tệp đã được trộn lẫn trong các tệp công khai.

Kết quả là dữ liệu bị tiếp xúc là 38 terabyte. Nó chứa mật khẩu của nhân viên, các khóa bí mật, các cuộc trò chuyện qua ứng dụng nhắn tin nội bộ và dữ liệu công việc riêng tư của hơn 35.000 nhân viên.

Một đội làm nghiên cứu AI đã mở công ty bên trong khi chia sẻ dữ liệu để đào tạo AI.

Amazon Q, AI dành cho doanh nghiệp của Amazon, đã hiển thị các lỗi rò rỉ dữ liệu khách hàng riêng tư cùng với ảo giác. Nền tảng AI nội bộ của nhóm tư vấn McKinsey đã trở thành mục tiêu của các hacker, và hàng triệu bản ghi khách hàng cũng như tài liệu dự án bí mật đã bị truy cập trong vòng hai giờ.

Đại lý AI của Meta cũng tiết lộ tài liệu nội bộ và thông tin người dùng. Một chương trình trợ lý AI đã để hàng nghìn máy chủ mở trên internet, bị tấn công bởi phần mềm độc hại và mất dữ liệu người dùng số lượng lớn.

Vào tháng 2 năm 2026, một phương pháp mới được tiết lộ thông qua nghiên cứu bảo mật. Một lời nhắc được tạo ra một cách độc hại có thể biến một cuộc hội thoại ChatGPT bình thường thành một kênh rò rỉ dữ liệu. Người dùng không biết gì. Lỗ hổng đã được sửa trong cùng tháng.

Đến thời điểm này là những sự cố. Những điều tiếp theo không phải là những sự cố.

Vào tháng 8 năm 2023, nền tảng hội thảo web Zoom đã âm thầm thay đổi các điều khoản sử dụng của nó. Nó đã cấp cho công ty quyền sử dụng video, âm thanh, ghi chép hội thoại và tệp được chia sẻ mà người dùng đã trao đổi cho việc đào tạo AI.

Hãy nhớ lại những gì mọi người đã làm trên Zoom trong thời gian làm việc tại nhà. Họ có các cuộc họp công ty, tham dự các lớp học, được tư vấn tại các bệnh viện và tham dự các dịp tang lễ.

Khi chỉ trích tăng lên, Zoom đã rút lại và nói rằng nó sẽ không sử dụng nó mà không có sự đồng ý.

Slack bị phát hiện vào năm 2024. Nó đã sử dụng các tin nhắn mà người dùng trao đổi và các tệp riêng tư mà họ tải lên để đào tạo AI theo mặc định. Để từ chối, người dùng phải gửi một yêu cầu riêng biệt và công ty không thông báo sự thật này đúng cách.

Giá trị mặc định quyết định mọi thứ. Hầu hết mọi người không thay đổi cài đặt. Các công ty biết sự thật này.

Adobe đã thay đổi điều khoản sử dụng đám mây để có thể phân tích các dự án bí mật và hình ảnh mà các nhà sáng tạo đang thực hiện nhằm mục đích đào tạo trí tuệ nhân tạo của công ty, và kết quả là đã phải đối mặt với một chiến dịch hủy đăng ký của các nhà thiết kế trên toàn thế giới.

LinkedIn đã sử dụng hồ sơ, kinh nghiệm làm việc và các bài đăng để đào tạo trí tuệ nhân tạo mà không có sự đồng ý trước. X đã thiết lập chế độ tự động thu thập tất cả các bài đăng để đào tạo Grok và đã bị các cơ quan chức năng của châu Âu điều tra. SoundCloud đã sử dụng các bài hát của các nhạc sĩ vào việc đào tạo.

Meta đã công bố sẽ sử dụng các bài viết và hình ảnh mà người dùng Facebook và Instagram đã đăng trong suốt hàng chục năm qua để đào tạo mô hình ngôn ngữ, và đã bị các cơ quan bảo vệ thông tin cá nhân của châu Âu và Nam Mỹ xử phạt.

Những bức ảnh em bé đăng từ mười năm trước, những bài viết đáng xấu hổ viết từ mười lăm năm trước, và những bức ảnh chụp cùng các thành viên trong gia đình đã qua đời đều nằm trong danh sách tài liệu đào tạo. Khi đăng lên, những thứ này vốn dĩ chỉ để cho bạn bè xem.

Người ta cũng phát hiện ra rằng Meta đã chuyển giao những bức ảnh và video riêng tư được chụp bằng kính thông minh Ray-Ban cho các nhân viên xử lý dữ

liệu thuê ngoài ở nước ngoài để sử dụng cho việc đào tạo. Điều này có nghĩa là một người nào đó đang ngồi trong một văn phòng ở một quốc gia nào đó đã gắn thẻ tên cho các đoạn video quay trong phòng khách của chúng ta.

Vào tháng 5 năm 2026, một vụ kiện tập thể đã được đệ trình lên Tòa án Liên bang California. Nội dung vụ kiện là do các công nghệ theo dõi như Meta Pixel và Google Analytics được gắn trên trang web ChatGPT, nên thông tin liên quan đến các cuộc hội thoại, ID người dùng và địa chỉ email đã bị truyền cho Google và Meta.

Họ lập luận rằng họ cứ tưởng chỉ có một đối tượng trò chuyện nhưng thực ra lại có tới ba.

Trong lĩnh vực hồ sơ y tế, quy mô lại hoàn toàn khác.

Vào năm 2017, Cơ quan Bảo vệ Dữ liệu Vương quốc Anh đã phán quyết rằng thỏa thuận chia sẻ dữ liệu giữa Bệnh viện Royal Free thuộc Dịch vụ Y tế Quốc gia và Google DeepMind đã vi phạm pháp luật. Bệnh viện này đã giao toàn bộ hồ sơ y tế của một triệu sáu trăm nghìn bệnh nhân mà không có sự đồng ý của họ, dưới danh nghĩa tạo ra một ứng dụng chẩn đoán tổn thương thận cấp tính.

Google sau đó đã ký hợp đồng với một trong những chuỗi bệnh viện lớn nhất nước Mỹ để nhận hồ sơ y tế cho mục đích đào tạo, và đã bị chỉ trích vì dự án bí mật thu thập dữ liệu y tế của năm mươi triệu người mà bệnh nhân không hề hay biết.

Trường hợp khiến chúng ta cảm thấy nặng lòng nhất chính là trường hợp cuối cùng này.

Crisis Text Line của Mỹ là một tổ chức phi lợi nhuận chuyên tư vấn phòng chống tự tử và khủng hoảng sức khỏe tâm thần. Khi có ý nghĩ muốn chết, nếu bạn gửi tin nhắn thì sẽ có tư vấn viên trả lời. Đây là nơi những người tìm kiếm sự giúp đỡ viết ra những lời nói tận đáy lòng mình.

Tổ chức này đã thu thập các hồ sơ trò chuyện đó và giao chúng cho một công ty con về trí tuệ nhân tạo vì mục đích lợi nhuận của họ. Công ty đó đã sử dụng dữ liệu này để đào tạo một chatbot dịch vụ khách hàng dành cho doanh nghiệp.

Những dòng tin nhắn muốn chết đã trở thành nguyên liệu để nâng cao chất lượng phục vụ khách hàng.

Ứng dụng sức khỏe Flo, nơi ghi lại chu kỳ kinh nguyệt và khả năng mang thai, đã bị Ủy ban Thương mại Liên bang xử phạt vì tự ý truyền dữ liệu sức khỏe của người dùng cho các công ty như Facebook và Google.

Khi đặt các sự việc này cạnh nhau, chúng ta sẽ thấy một thực tế là luồng dữ liệu chỉ chảy theo một hướng.

Dữ liệu đi từ con người đến công ty. Thứ quay trở lại từ công ty đến với con người là các tính năng tiện lợi. Sự tiện lợi đó là có thật. Vì vậy, giao dịch này mới được thành lập.

Vấn đề nằm ở chỗ nó không có bảng giá. Hợp đồng không ghi rõ chúng ta đưa cái gì và nhận lại cái gì. Ngay cả khi có ghi, thì điều đó cũng nằm ở trang thứ mười hai của một bản điều khoản dài mười tám trang.

Và dữ liệu một khi đã được chuyển đi thì sẽ không quay trở lại. Cho dù công ty có nói là đã xóa, dữ liệu đó cũng không thể bị xóa bên trong mô hình đã được đào tạo bằng chính dữ liệu đó.

Mô hình không biết cách để quên đi những gì nó đã học.

3.3

Theo dõi vị trí và giám sát người lao động quá mức

Amazon đã bộc lộ một bằng sáng chế. Đó là một chiếc vòng đeo tay.

Khi một công nhân làm việc ở kho hàng đeo chiếc vòng này, tay của họ sẽ được theo dõi theo thời gian thực xem nó đi đến kệ nào, hộp nào. Nếu tay đi đến một nơi không đúng, vòng sẽ rung.

Nó báo hiệu một cách nhẹ nhàng. Không phải chỗ đó, mà là chỗ này.

Tôi không phải là người duy nhất liên tưởng đến chiếc vòng huấn luyện chó.

Trong thiết kế của thiết bị này có một quan điểm cụ thể về con người. Quan điểm rằng công nhân là thiết bị đầu cuối của hệ thống, tay là cái kẹp của thiết bị đó, và rung là phản hồi để giảm sai số.

Công ty con của Amazon ở Pháp thực sự bị phạt. Mức phạt là ba mươi hai triệu euro.

Lý do là như thế này. Khi công nhân kho hàng cầm máy quét mã vạch làm việc, công ty đo khoảng thời gian giữa mỗi lần quét bằng từng giây. Nếu không có lần quét nào trong hơn mười phút, cảnh báo sẽ tích lũy.

Mười phút. Nếu đi vệ sinh về cũng mất mười phút. Nếu uống nước và thở dài cũng mất mười phút.

Cơ quan Bảo vệ Dữ liệu Pháp cho rằng sự giám sát này là quá mức.

Trước khi đo lường công nhân, các công ty đã học cách bán vị trí trước.

Năm 2021, một công ty tên là Life360 bị đưa vào tầm ngắm. Đó là một ứng dụng để các gia đình kiểm tra lẫn nhau ở đâu. Cha mẹ cài đặt nó để xem con cái có

đến trường an toàn không.

Công ty này đã bán hồ sơ quỹ đạo chuyển động và địa điểm thăm viếng của hàng chục triệu người dùng cho các công ty trung gian dữ liệu.

Dữ liệu vị trí có bản chất khác với thông tin khác. Nếu biết ai đi đâu, bạn gần như biết tất cả về người đó. Vào tối thứ năm hàng tuần đi đến tòa nhà nào, đi đến bệnh viện nào, đi đến cơ sở tôn giáo nào, đêm nào không ở nhà.

Xóa tên cũng không có ích. Nơi ở mỗi đêm là nhà, nơi ở mỗi ngày là nơi làm việc. Chỉ cần hai điểm thôi là có thể xác định được con người.

Công ty trung gian dữ liệu vị trí Tamoco bị phát hiện thu thập hồ sơ GPS của công dân Na Uy và phân phối chúng cho cơ quan điều tra và quân sự. Công ty Anh Quốc Hook cũng bị xử phạt vì bán dữ liệu vị trí được thu thập qua ứng dụng cho bên thứ ba.

Nền tảng giao hàng Trung Quốc Meituan bị phản đối vì theo dõi quá mức theo thời gian thực cả tài xế giao hàng và người dùng.

Cơ quan điều tra Hoa Kỳ đã sử dụng hệ thống camera nhận diện biển số tự động để theo dõi quỹ đạo chuyển động của những người tham gia biểu tình và các nhà hoạt động mà không có lệnh của tòa án, đưa họ vào mạng lưới giám sát.

Không thể che biển số. Pháp luật yêu cầu phải gắn nó.

Ở những địa điểm giao hàng và vận chuyển, camera nhìn thấy khuôn mặt.

Xe giao hàng thuê của Amazon được lắp camera giám sát video nhân tạo thông minh và chương trình đánh giá tài xế. Nó nhìn thấy hướng nhìn, ngáp hay không, cơ mặt chuyển động như thế nào, bước phanh mạnh bao nhiêu và bao nhiêu đột ngột.

Vấn đề là hệ thống này lại phạt lái xe an toàn.

Quay vô lăng để tránh cành cây bị bắt là bất cẩn. Nhìn gương chiếu hậu bị bắt là mất tập trung. Đó là những hành động mà trường dạy lái xe dạy phải làm.

Nếu điểm số bị trừ, trợ cấp sẽ bị trừ và lượng công việc được phân công sẽ giảm. Tài xế phải lựa chọn giữa lái xe an toàn và điểm số.

Thuật toán Amazon Flex yêu cầu con đường ngắn nhất về mặt toán học mà không tính đến công trình đường bộ hoặc thời tiết xấu. Những tài xế đã đi vào những con đường nguy hiểm và bùn lầy.

Đường trên bản đồ là ngắn. Nhưng bản đồ không cho thấy cái gì ở trên đường đó.

Văn phòng cũng không phải là nơi an toàn.

Công ty tài chính Barclays đã gắn cảm biến dưới bàn nhân viên tại trụ sở ở Anh và cài đặt trình giám sát trên máy tính làm việc. Nó đo thời gian ngồi tại chỗ và tần suất gõ bàn phím bằng từng giây để tính tổng thời gian không năng suất.

Khi công đoàn phản đối, công ty dừng hoạt động.

Microsoft đã phát hành một tính năng được gọi là điểm số năng suất. Nó gán điểm cho từng cá nhân dựa trên số lần gửi email, tốc độ trả lời trên ứng dụng nhắn tin, mức độ tham gia hội nghị, và hiển thị cho quản lý viên.

Những chỉ số này có điểm chung. Chúng chỉ đếm những thứ dễ đếm.

Giữa người nhìn ra cửa sổ và suy tư một giờ để giải quyết vấn đề, và người trả lời tin nhắn hai mươi ba lần trong một giờ, ai làm việc hơn? Điểm số chọn người sau.

Vì thế mọi người bắt đầu làm điều đó. Điểm số thay đổi hành động, và hành động thay đổi lại tạo ra điểm số.

Meta cũng bị phát hiện cài đặt phần mềm trên máy tính công việc của những nhân viên Hoa Kỳ để thu thập chuyển động chuột, nhấp chuột, nhập bàn phím,

và ảnh chụp màn hình theo thời gian thực.

Thậm chí còn có hệ thống đo cảm xúc.

Công ty trung tâm cuộc gọi toàn cầu Teleperformance đã triển khai một hệ thống bật webcam của những công nhân làm việc tại nhà suốt giờ làm việc. Trí tuệ nhân tạo theo dõi góc khuôn mặt và vị trí của con mắt để phát hiện liệu mắt rời khỏi màn hình, xem điện thoại, hay có người khác đi vào phòng, và thông báo cho công ty.

Phòng khách nhà của bạn được quay phim suốt giờ làm việc. Nếu gia đình đi qua, cảnh báo sẽ phát ra.

Tại một văn phòng Canon ở Trung Quốc, họ sử dụng một hệ thống yêu cầu phải cười trước camera để cửa mở. Phải nhận diện được nụ cười thì mới xác nhận được thời gian đến.

Công ty nói rằng mục đích là tạo ra một bầu không khí sáng sủa.

Ai cũng biết nụ cười phía trước cánh cửa yêu cầu phải cười là kiểu nụ cười nào.

Công ty kế toán PricewaterhouseCoopers gây tranh cãi khi triển khai một công cụ sử dụng nhận diện khuôn mặt để theo dõi ngay cả thời gian nhân viên giao dịch viên tài chính đi vệ sinh.

Một nhà sản xuất Tây Ban Nha tên là Plastic Forte bị phạt vì sử dụng nhận diện khuôn mặt để giám sát thời gian đến và đi, cũng như chuyển động của công nhân dòng chuyền. Một công ty dịch vụ Anh Quốc tên là Serco cũng nhận được lệnh dừng sử dụng sau khi sử dụng nhận diện khuôn mặt một cách bất hợp pháp để giám sát nhân viên.

Có những nỗ lực đo tông giọng, nhịp tim, nhiệt độ cơ thể, thậm chí phản ứng điện da da để theo dõi trạng thái cảm xúc. Một số hệ thống đặt camera khắp sàn nhà máy để nhìn thấy quỹ đạo chuyển động và động tác của công nhân.

Những con số được thu thập như thế này cuối cùng sẽ cắt giảm con người.

Năm 2021, công ty nền tảng trò chơi Xsolla đã sa thải một trăm năm mươi nhân viên cùng một lúc dựa trên kết quả phân tích thuật toán giám sát.

Tiêu chí như thế này. Lịch sử sử dụng chương trình nội bộ, tần suất lưu mã và nhập tài liệu, mức độ tham gia phòng trò chuyện và làm việc tài liệu. Những người có điểm thấp trên những chỉ số này đã được phân loại là nhân viên không hoạt động.

Những nhân viên không biết họ được đánh giá theo tiêu chí nào. Họ không biết ngữ cảnh công việc nào bị thiếu. Họ đã rời khỏi một công ty nơi họ làm việc lâu năm bằng một thông báo tự động duy nhất.

Người nói chuyện điện thoại lâu với khách hàng, người giữ hậu bối lại để hướng dẫn, người giải quyết vấn đề bằng lời nói trong phòng họp. Hãy thử nghĩ xem những người này được ghi chép vào ô nào. Họ không được ghi vào bất kỳ ô nào cả.

Thuật toán dự đoán nghỉ việc của một công ty đã theo dõi lịch sử tìm kiếm và hoạt động trên nền tảng của nhân viên để tìm ra trước những người có vẻ sẽ nghỉ việc và gây bất lợi cho họ.

Nếu bị đối xử tệ vì dự đoán có vẻ sẽ rời đi, họ sẽ thực sự rời đi. Dự đoán đó là đúng.

Trong công việc trên nền tảng, ngay cả việc sa thải cũng là tự động.

Các tài xế giao hàng của Amazon Flex đã bị sa thải bằng một email tự động mà không hề gặp mặt một người nào. Dù ứng dụng bị lỗi, hàng ra khỏi trung tâm kho vận trễ, kẹt xe hay mưa lớn đổ xuống, tất cả đều bị tính vào điểm giao hàng chậm trễ. Khi điểm bị tích luỹ, tài khoản sẽ bị đình chỉ vĩnh viễn.

Nếu họ phản đối, một email trả lời được viết sẵn sẽ được gửi đến.

Thuật toán Frank của Deliveroo mà Tòa án Bologna ở Ý coi là có vấn đề đã xuất hiện trước đó. Dù bị ốm không thể đi làm, có tang gia hay tham gia vào cuộc

đình công được pháp luật bảo đảm, điểm số vẫn bị trừ.

Pháp luật trao quyền đình công. Thuật toán khiến bạn phải chịu đói nếu sử dụng quyền đó. Cả hai cùng hoạt động song song trong cùng một quốc gia.

Cũng có những người bị sa thải vì hệ thống không nhận ra khuôn mặt.

Pa Edrissa Manjang và Imran Raza, những người giao hàng cho Uber Eats ở Anh, đã bị sa thải chỉ sau một đêm do phần mềm xác nhận khuôn mặt tự động của nền tảng này.

Hệ thống không nhận diện tốt khuôn mặt của những nhân viên giao hàng có màu da tối. Nếu nhận diện thất bại, hệ thống sẽ đánh giá họ là người dùng gian lận, cho người khác mượn tài khoản. Tài khoản bị đóng băng ngay lập tức.

Hai người này đã làm việc nhiều năm với những đánh giá tốt. Bản thân họ phải trả giá cho việc máy móc không nhận ra khuôn mặt của họ.

Ở nhiều nơi bao gồm bang Utah và New South Wales, tài khoản của các tài xế chuyển giới có khuôn mặt thay đổi sau phẫu thuật chuyển giới hoặc điều trị hormone đã bị đình chỉ hàng loạt. Hệ thống đã không tính đến sự thật là con người có thể thay đổi.

Các nền tảng giao hàng như Instacart và Shipt cũng cắt giảm tiền lương và tự động sa thải người lao động bằng một thuật toán không minh bạch. Thuật toán tính lương người giao hàng của Bưu chính Hoa Kỳ cũng cắt giảm phụ cấp mà không công khai công thức.

Trường hợp cay đắng nhất là phần cuối cùng này.

Vào năm 2025 tại một ngân hàng lớn của Úc, các nhân viên đã dạy cho chatbot dịch vụ khách hàng trong vài tháng. Họ nhập bí quyết ứng xử mà mình tích lũy được, hướng dẫn cách trả lời trong những tình huống khó khăn và sửa chữa khi chatbot làm sai.

Khi hiệu suất của chatbot tăng lên, công ty đã thông báo sa thải tập thể đối với những nhân viên đó.

Họ đã tự tay dạy cho chính người thay thế mình.

Hãy thử liệt kê các công cụ đã xuất hiện trong phần này. Vòng đeo tay, đồng hồ bấm giờ máy quét mã vạch, camera xe hơi, cảm biến bàn làm việc, máy ghi bàn phím, webcam, cửa nhận diện nụ cười, camera theo dõi nhà vệ sinh, mô hình dự đoán nghỉ việc.

Tất cả đều hướng vào con người. Và tất cả đều được gọi bằng cùng một cái tên. Đó là quản lý năng suất.

Người lao động không thể xem điểm số của mình. Họ cũng không biết nó được tính toán như thế nào. Nếu hỏi, họ được trả lời đó là bí mật kinh doanh.

Có một điều chắc chắn. Những hệ thống này rất thành thạo trong việc tìm ra thời gian nghỉ ngơi của con người. Vì thời gian nghỉ ngơi rất dễ đếm.

Khoảnh khắc làm việc hiệu quả rất khó đếm. Vì vậy chúng không được đếm.

CHƯƠNG 4

Bản quyền và quyền của người sáng tạo: AI tạo sinh và cuộc tranh chấp về dữ liệu huấn luyện

4.1

Các vụ kiện vi phạm bản quyền và cạo dữ liệu trái phép

Đầu tiên, họ mua sách. Đó là sách thật. Họ trả tiền để mua nó.

Tiếp theo, họ cắt bỏ gáy sách. Nếu dùng dao cắt bỏ gáy, cuốn sách sẽ trở thành một tập giấy rời. Họ đưa những tờ giấy rời này vào máy quét nạp giấy tự động. Từng tờ giấy chạy qua và trở thành tệp kỹ thuật số.

Những tờ giấy sau khi quét xong sẽ bị vứt bỏ.

Bên trong công ty trí tuệ nhân tạo Anthropic, có hàng trăm ngàn cuốn sách đã được xử lý như vậy. Việc này mang tên Project Panama.

Về mặt pháp lý, đây là một phương pháp thông minh. Sách đã mua là tài sản sở hữu nên có thể cắt bỏ. Tuy nhiên, nếu hình dung cảnh tượng này trong đầu, chúng ta sẽ thấy rất kỳ lạ. Một cuốn sách mà ai đó dành cả đời để viết bị xử lý trên băng chuyền chỉ trong hai mươi giây rồi đi vào thùng rác.

Anthropic không chỉ sử dụng phương pháp này. Họ còn tải xuống miễn phí hơn năm trăm ngàn cuốn sách từ các thư viện lậu như LibGen và Pirimi.

Trong một vụ kiện tập thể do các tác giả Andrea Bartz, Kirk Wallace Johnson và Charles Graeber đệ trình, vào năm 2025 Anthropic đã đồng ý bồi thường một tỷ năm trăm triệu đô la. Đây là số tiền lớn nhất trong lịch sử các vụ kiện bản quyền tại Mỹ.

Điều quan trọng trong vụ án này là ranh giới mà tòa án đã vạch ra. Tòa án không coi bản thân hành vi sử dụng các tác phẩm có bản quyền để huấn luyện mô hình ngôn ngữ là bất hợp pháp. Vấn đề bị lên án là phần tải xuống và lưu trữ các bản lậu.

Điều này có nghĩa là học hỏi thì được, nhưng ăn cắp để học thì không. Sự phân biệt này sẽ trở thành nền tảng cho tất cả các vụ kiện trong tương lai.

Các tác giả bắt đầu lên tiếng lần đầu tiên là vào năm 2023.

Vào tháng 6 năm đó, các tiểu thuyết gia Mona Awad và Paul Tremblay đã đệ trình vụ kiện đầu tiên chống lại OpenAI. Vào tháng 9, mười bảy thành viên của Hiệp hội Tác giả Mỹ đã nộp đơn kiện tập thể lên Tòa án Liên bang New York. Đó là những cái tên như George Martin người viết Trò chơi vương quyền, John Grisham, Jodi Picoult, Scott Turow, David Baldacci và Jonathan Franzen.

Lập luận của họ như sau: OpenAI đã sử dụng hàng chục ngàn cuốn tiểu thuyết để huấn luyện mà không xin phép. Nguồn nguyên liệu đó đến từ các thư viện lậu như Library Genesis và Z-Library, cùng một tập dữ liệu mang tên Books3.

Lý do ChatGPT viết ra những câu văn trôi chảy là vì nó đã đọc rất nhiều câu văn trôi chảy. Những câu văn đó đã được ai đó viết ra.

Người từng đoạt giải Pulitzer Michael Chabon, nhà viết kịch David Henry Hwang, và tác giả sách phi hư cấu Rachel Louise Snyder cũng đã đứng lên chống lại OpenAI và Meta.

Vào tháng 11 năm 2023, tác giả Julian Sancton đã đệ trình một vụ kiện tập thể chống lại OpenAI và Microsoft, cáo buộc họ sử dụng trái phép hàng chục ngàn cuốn sách phi hư cấu. Các tác giả sách tôn giáo, bao gồm cả Mike Huckabee, đã khởi kiện Meta và Microsoft, cho rằng một trăm tám mươi ngàn cuốn sách trong tập dữ liệu Books3 đã được sử dụng để huấn luyện.

Vào tháng 8 năm 2024, các nhà báo Nicholas Basbanes và Nicholas Gage đã tham gia vụ kiện, nêu vấn đề về một tập dữ liệu khác mang tên Books2.

Vào tháng 10 năm 2025, giáo sư Susana Martinez-Conde và giáo sư Stephen Macknik của Trung tâm Khoa học Y tế Downstate thuộc Đại học Bang New York đã nộp đơn kiện Apple. Họ cáo buộc rằng dữ liệu từ các thư viện lậu đã được sử dụng để huấn luyện Apple Intelligence. Nvidia cũng bị ba tác giả khởi kiện vì lý

do tương tự.

Các công ty truyền thông đang đấu tranh từ một góc độ khác.

Vào tháng 12 năm 2023, New York Times đã nộp đơn kiện trị giá hàng tỷ đô la chống lại OpenAI và Microsoft. Họ cho rằng hàng triệu bài báo đã được sử dụng để huấn luyện mà không có sự cho phép hay đền bù nào.

Trọng tâm của vụ kiện này không phải là việc huấn luyện, mà là sự cạnh tranh. Một dịch vụ được tạo ra từ kết quả thu thập tin tức mà tòa soạn phải bỏ tiền bạc và thời gian ra, nay lại cạnh tranh giành độc giả với chính tòa soạn đó. Nếu độc giả chỉ đọc phần tóm tắt và không truy cập vào bài báo, tòa soạn sẽ mất đi doanh thu quảng cáo.

Việc lấy tin đòi hỏi tiền bạc. Phải cử phóng viên đến hiện trường, yêu cầu cung cấp tài liệu và gặp gỡ nguồn tin nhiều lần. Việc tóm tắt không tốn những chi phí đó.

Vào năm 2026, cuộc chiến này vẫn đang tiếp diễn. New York Times và nhiều công ty truyền thông khác đã yêu cầu thẩm phán liên bang ở Manhattan đưa ra các biện pháp trừng phạt, cho rằng OpenAI đã lừa dối tòa án về khả năng tìm kiếm các tài liệu có bản quyền trong hệ thống của công ty.

Vào tháng 4 năm 2024, tám tờ nhật báo của Mỹ, bao gồm Chicago Tribune và Denver Post, đã nộp đơn kiện. Trong đó có cáo buộc rằng họ không chỉ thu thập bài báo trái phép mà còn cố tình xóa bỏ các ký hiệu bản quyền.

Hành vi xóa bỏ nguồn gốc khó có thể coi là một sai sót. Phải viết mã như vậy trong mã nguồn thì nó mới bị xóa.

Vào tháng 2 năm 2024, The Intercept, Raw Story và AlterNet đã đệ đơn kiện, và vào tháng 6, Trung tâm Báo chí Điều tra cũng khởi kiện. Vào tháng 11, các nhà xuất bản Canada sở hữu tờ Calgary Herald và Montreal Gazette đã bắt đầu hành động.

Dịch vụ tìm kiếm trí tuệ nhân tạo Perplexity cũng trở thành mục tiêu. Vào tháng 10 năm 2024, Dow Jones và New York Post đã khởi kiện, và vào tháng 12 năm 2025 là Chicago Tribune. Condé Nast và New York Times đã gửi thư cảnh báo.

Chính quyền Pháp đã phạt Google hai trăm năm mươi triệu euro vào tháng 3 năm 2024. Lý do là vì họ đã sử dụng các bài báo tin tức để huấn luyện Gemini mà không bàn bạc với các công ty truyền thông.

Tại Hàn Quốc, vào năm 2024 và 2025, các đài phát thanh truyền hình cũng đã nộp đơn kiện chống lại Naver. Cáo buộc cho rằng nội dung tin tức phát sóng đã bị sử dụng trái phép để huấn luyện HyperCLOVA X.

Về phía hình ảnh, bằng chứng đã in hằn rõ ràng trên màn hình.

Vào tháng 1 năm 2023, Getty Images đã đệ trình vụ kiện chống lại Stability AI tại Anh và Mỹ. Họ cho rằng hơn mười hai triệu bức ảnh có đóng dấu mờ của công ty đã được sử dụng để huấn luyện mà không có giấy phép.

Bằng chứng là thế này. Ở một góc của bức tranh do Stable Diffusion tạo ra, dấu mờ của Getty Images vẫn còn lưu lại dưới dạng bị bóp méo.

Cỗ máy không biết dấu mờ là gì. Vì đó là họa tiết thường xuyên xuất hiện trên các bức ảnh nên nó đã học luôn cả phần đó. Điều này chẳng khác nào một tên trộm mang món đồ ăn cắp ra ngoài mà vẫn để nguyên bảng tên của cửa hàng.

Nhiếp ảnh gia kho ảnh người Đức Robert Kneschke đã xác nhận rằng ảnh của mình bị sử dụng để huấn luyện thông qua tập dữ liệu LAION và đã tiến hành khởi kiện. Họa sĩ minh họa Hollie Mengert đã phải trải qua việc toàn bộ phong cách vẽ của mình bị sao chép và trở thành một mô hình.

Rất khó để bảo vệ phong cách vẽ bằng bản quyền. Bởi vì những gì luật pháp bảo vệ là các tác phẩm cá biệt chứ không phải phong cách. Thế nhưng, thứ mà trí tuệ nhân tạo đánh cắp lại chính xác là phong cách.

Có một sự sai lệch giữa ranh giới do luật pháp vạch ra và những gì công nghệ đang lấy đi.

Các tập đoàn truyền thông khổng lồ cũng đã bắt đầu vào cuộc.

Năm 2025, Disney, Universal và Warner Bros. Discovery đã kiện công ty tạo ảnh Midjourney. Lý do là các nhân vật trong hoạt hình và phim của họ đã bị thu thập trái phép và xuất hiện dưới dạng ảnh tổng hợp. Các công ty này cũng đã nhắm vào Minimax của Trung Quốc.

Công ty phim Alcon Entertainment đã kiện Tesla và Elon Musk vào tháng 10 năm 2024. Lý do là tại sự kiện công khai xe tự lái, họ đã sử dụng hình ảnh thị giác từ Blade Runner 2049 được tái tạo bằng trí tuệ nhân tạo.

Năm 2024, tòa án Trung Quốc đã ra phán quyết vi phạm bản quyền đối với một công ty đã huấn luyện trái phép nhân vật Ultraman.

Âm nhạc tồn tại dưới dạng âm thanh.

Tháng 6 năm 2024, Sony Music, Universal Music Group và Warner Music đã kiện các công ty tạo nhạc Suno và Udio. Họ lập luận rằng những công ty này đã cạo trái phép các bản nhạc thương mại tích lũy trong hàng chục năm để huấn luyện, dẫn đến một cỗ máy có thể sao chép một cách tinh vi tông giọng và phong cách của một ca sĩ cụ thể.

Tháng 10 năm 2023, Universal Music Group đã kiện Anthropic về việc thu thập trái phép cơ sở dữ liệu lời bài hát. Lý do là Claude đã in ra nguyên văn các lời bài hát nổi tiếng. Trong vụ kiện này, một sự cố đã được đề cập ở trước nơi các tài liệu do Anthropic nộp đã trích dẫn một bài báo học thuật không tồn tại.

Diễn viên hài George Carlin qua đời vào năm 2008. Tháng 1 năm 2024, gia đình của anh ấy đã kiện một kênh đã tạo ra một chương trình đặc biệt hài kịch AI bằng cách huấn luyện giọng nói của Carlin khi còn sống, dẫn đến việc xóa và giải quyết.

Người chết không nói những điều cười mới. Nhưng giọng nói của họ vẫn nói những điều cười mới.

Câu chuyện về những người sống từ giọng nói của họ tìm trực tiếp hơn.

Diễn viên lồng tiếng lão làng Paul Sky Lieman và Linea Sage đã kiện một công ty khởi động tạo giọng nói tại tòa án liên bang New York vào tháng 5 năm 2024. Họ lập luận rằng công ty đã ghi âm giọng nói của họ dưới danh nghĩa mục đích nghiên cứu hội thoại, sau đó sử dụng các bản ghi âm đó làm dữ liệu huấn luyện cho dịch vụ diễn viên lồng tiếng AI thương mại và bán nó.

Diễn viên lồng tiếng Bev Standing nhận ra rằng giọng nói của cô ấy đã được sử dụng làm giọng đọc tự động của TikTok và đã kiện để nhận được thỏa thuận. Hàng trăm triệu người đã nghe những lời nói mà cô ấy không nói bằng giọng nói của cô ấy.

Diễn viên Scarlett Johansson đã phản đối vào tháng 5 năm 2024 rằng một giọng nói AI do OpenAI tạo ra đã bắt chước giọng nói của cô ấy. OpenAI đã nhanh chóng ngừng dịch vụ âm thanh đó.

Johansson cũng bắt đầu hành động pháp lý chống lại một nhà phát triển ứng dụng đã tạo quảng cáo AI mà cô ấy xuất hiện mà không được phép. Diễn viên Bryan Cranston cũng tiết lộ rằng giọng nói và khuôn mặt của anh ấy đã được sử dụng để huấn luyện mô hình video mà không có sự đồng ý.

Thậm chí cả cơ sở dữ liệu chuyên môn cũng bị cạo.

Tháng 11 năm 2024, cơ sở dữ liệu luật pháp Canada CanLii đã kiện một công ty khởi động AI pháp luật. Họ đã cạo hàng chục ngàn bản án và tài liệu giải thích mà không quan tâm đến các điều khoản dịch vụ và hạn chế quyền truy cập.

Công ty thông tin pháp lý Thomson Reuters đã thắng kiện chống lại Ross Intelligence vào tháng 1 năm 2025 về việc nó đã huấn luyện trái phép cơ sở dữ liệu của nó.

Luận điểm phòng vệ mà các công ty đưa ra trong những vụ kiện này gần như là như nhau. Đó là công bằng sử dụng. Huấn luyện không phải là sao chép mà là phân tích, và nó không khác gì khi con người đọc một cuốn sách và học tập.

Có một điều còn thiếu trong phép so sánh này. Con người đọc từng cuốn sách một, phải mất thời gian để đọc, và họ mua hoặc mượn những cuốn sách họ đã đọc. Và khả năng mà một người đạt được từ việc đọc chỉ thuộc về người đó.

Mô hình đọc năm trăm ngàn cuốn sách trong vài ngày và bán khả năng đó đồng thời cho hàng trăm triệu người.

Khi quy mô thay đổi, bản chất cũng thay đổi. Múc một cốc nước và kéo một con sông là những hành động khác nhau.

Một điều đã rõ ràng. Bất kể các vụ kiện này kết thúc như thế nào, vẫn chưa có cách nào để xóa các dấu vết của một tác giả cụ thể từ một mô hình đã được huấn luyện.

Các cuốn sách đã bị cắt, quét xong rồi, giấy đã bị vứt.

4.2

Sự phủ nhận bản quyền đối với sản phẩm do AI tạo ra và sự hỗn loạn của hệ sinh thái nghệ thuật

Chris Kashtanova đã tạo ra một cuốn tiểu thuyết đồ họa vào mùa thu năm 2022. Tiêu đề của nó là Zarya of the Dawn.

Tác giả đã tự viết câu chuyện. Hình ảnh được tạo ra bằng cách nhập câu lệnh vào Midjourney. Tác giả đã liên tục chỉnh sửa câu lệnh cho đến khi có được bức tranh ưng ý.

Kashtanova đã nộp đơn đăng ký lên Cục Bản quyền Hoa Kỳ và đã được chấp thuận.

Tuy nhiên, khi sự thật rằng các bức tranh được tạo ra bởi trí tuệ nhân tạo bị lộ, Cục Bản quyền đã xem xét lại. Quyết định được đưa ra vào tháng 2 năm 2023.

Văn bản do chính tác giả viết và cách sắp xếp văn bản được bảo vệ bản quyền. Việc đăng ký cho các bức tranh do Midjourney tạo ra bị hủy bỏ.

Lý lẽ của Cục Bản quyền như sau. Dù bạn có viết câu lệnh cẩn thận đến đâu, bạn không thể biết trước kết quả sẽ ra sao và cũng không thể điều chỉnh chi tiết được. Điều này khác với việc một họa sĩ vẽ một đường bằng cọ.

Ý nghĩa của quyết định này rất quan trọng. Bức tranh do trí tuệ nhân tạo tạo ra không thuộc về ai cả. Ngay cả người tạo ra nó cũng không được sở hữu. Bất cứ ai cũng có thể lấy và sử dụng nó.

Lý do các công ty sử dụng hình ảnh trí tuệ nhân tạo thường là vì chi phí. Tuy nhiên, các trang bìa hoặc hình ảnh quảng cáo được tạo ra theo cách đó không thuộc sở hữu hợp pháp của họ. Không có căn cứ nào phù hợp để ngăn cản nếu đối thủ cạnh tranh lấy và sử dụng nguyên xi.

Họ đã làm ra nó với giá rẻ nhưng lại không thể bảo vệ nó.

Trước đây cũng đã từng có tranh cãi về việc tác giả là ai.

Tại phiên đấu giá Christie's ở New York vào năm 2018, một bức chân dung do trí tuệ nhân tạo vẽ đã được bán với giá bốn trăm ba mươi hai nghìn năm trăm đô la. Bức tranh có tiêu đề là Edmond de Belamy, và được đưa ra bởi nhóm nghệ thuật Pháp Obvious.

Có một sự thật đã được tiết lộ sau đó. Mã nguồn để tạo ra bức tranh đã được một nhà phát triển khác công khai. Vậy thì tác giả của bức tranh này là người đã nhập câu lệnh, người đã viết mã nguồn, hay là những họa sĩ ngày xưa đã vẽ những bức chân dung được dùng để học hỏi?

Trong danh mục đấu giá, công thức thuật toán đã được viết ở vị trí chữ ký. Thế nhưng tiền trúng đấu giá lại do con người nhận.

Khi bảo tàng Mauritshuis ở Hà Lan trưng bày bản sao do trí tuệ nhân tạo của bức tranh Cô gái đeo hoa tai ngọc trai, những lời chỉ trích cũng đã nổ ra. Đó là bởi vì họ đã treo một bản sao của máy móc vào vị trí của tác phẩm gốc.

Đến đây là câu chuyện về luật pháp và các bảo tàng nghệ thuật. Cuộc chiến thực sự đã nổ ra trên các bìa sách.

Nhà xuất bản Tor Books đã mua và sử dụng hình ảnh trí tuệ nhân tạo cho bìa tác phẩm mới của tác giả sách bán chạy Christopher Paolini. Khi sự thật được tiết lộ, độc giả và các họa sĩ minh họa đã phản đối.

Nhà xuất bản Bloomsbury cũng đã nhận chỉ trích sau khi sử dụng hình ảnh trí tuệ nhân tạo cho bìa tiểu thuyết của tác giả nổi tiếng Sarah J. Maas. DC Comics đã tung ra các bìa thay thế được làm bằng trí tuệ nhân tạo, nhưng đã phải thu hồi chúng do sự phản đối của các tác giả trực thuộc.

Vẽ tranh bìa là một công việc lớn đối với họa sĩ minh họa. Một bức tranh mất vài tuần để hoàn thành, và họ sống trong nhiều tháng bằng thu nhập đó. Nếu trí

tuệ nhân tạo đảm nhận phần bìa sách, những công việc đó sẽ biến mất.

Và trí tuệ nhân tạo đó đã học hỏi từ chính những bức tranh của các họa sĩ minh họa ấy.

Sự việc ngượng ngùng nhất đã đến từ một công ty sản xuất bảng vẽ.

Wacom tạo ra bảng vẽ đồ họa dành cho những người vẽ tranh. Khách hàng của họ chính là các họa sĩ minh họa. Nhân dịp năm con Rồng 2024, khi thực hiện quảng cáo cho thị trường Mỹ, họ đã sử dụng một hình ảnh con rồng do trí tuệ nhân tạo tạo ra.

Họ đã quảng cáo bằng một bức tranh thay thế công việc của chính những người đang mua sản phẩm của họ.

Công ty đã giải thích rằng đó là hình ảnh do một công ty quảng cáo thuê ngoài làm ra và lên tiếng xin lỗi. Trong cộng đồng những người vẽ tranh, cái tên của công ty này đã bị gọi chệch đi trong một thời gian.

Liên hoan Văn học Bradford đã sử dụng hình ảnh trí tuệ nhân tạo trong tài liệu quảng bá, khiến một họa sĩ minh họa dự kiến tham dự đã từ chối tham gia.

Trong ngành công nghiệp trò chơi, lời hứa đã bị phá vỡ hai lần.

Wizards of the Coast, công ty sản xuất Dungeons & Dragons, đã bị chỉ trích khi việc họ đưa bức tranh được tạo ra bằng công cụ trí tuệ nhân tạo vào minh họa cho cuốn sách mới bị phát hiện. Công ty đã thay đổi chính sách để không sử dụng hình ảnh trí tuệ nhân tạo nữa.

Thế nhưng họ lại tiếp tục sử dụng tranh trí tuệ nhân tạo trong các ấn phẩm quảng bá ngay sau đó.

Activision, công ty tạo ra Call of Duty, đã bán các vật phẩm trang trí trong trò chơi bằng trí tuệ nhân tạo để thu phí. Lego đã nhận phải sự phản đối sau khi dùng trí tuệ nhân tạo để tổng hợp một nhân vật không có bản quyền và sử dụng trong ấn phẩm quảng bá.

Trò chơi Palworld, có nét tương đồng với Pokémon, đã vướng vào nghi ngờ sử dụng trí tuệ nhân tạo để sao chép hình dáng nhân vật.

Lựa chọn của những người đứng đầu các công ty công nghệ cũng trở thành đề tài bàn tán.

Mark Zuckerberg của Meta đã làm một cuốn truyện tranh thiếu nhi cho các con gái nhỏ của mình, yêu cầu phần minh họa được vẽ bằng trình tạo hình ảnh trí tuệ nhân tạo của công ty và công bố nó.

Một trong những người giàu nhất thế giới đã không giao việc vẽ tranh cho sách của con mình cho con người. Chắc hẳn không phải vì ông muốn tiết kiệm tiền vẽ tranh. Có lẽ đơn giản chỉ vì làm như vậy sẽ nhanh hơn.

Những khoảnh khắc chọn cách làm nhanh chóng tích tụ lại sẽ làm thay đổi cả ngành công nghiệp.

Amazon đã tổng hợp và tạo ra poster quảng bá cho bộ phim truyền hình Fallout bằng trí tuệ nhân tạo, sau đó nhận phải chỉ trích vì sự thô sơ của nó.

Trong lĩnh vực phim ảnh, những lỗi sai đã bị in ra y nguyên.

A24, công ty sản xuất bộ phim Civil War, đã tạo ra và công bố poster quảng bá mô tả cảnh các thành phố lớn ở Mỹ bị phá hủy bằng trí tuệ nhân tạo. Các địa danh trong poster bị sai lệch vị trí, và cơ thể con người có những khớp xương bị bẻ gập một cách kỳ lạ.

Hơn nữa, những cảnh đó không hề xuất hiện trong phim.

Netflix đã xử lý hình nền trong bộ phim hoạt hình The Dog & The Boy bằng trí tuệ nhân tạo và ghi vào phần giới thiệu là nhà phát triển trí tuệ nhân tạo. Vẽ bối cảnh là nơi những người mới trong ngành công nghiệp hoạt hình trau dồi kỹ năng của họ. Nếu vị trí đó biến mất, thế hệ tiếp theo sẽ không thể phát triển.

Poster quảng bá cho phần 2 của bộ phim hoạt hình Arcane cũng được tạo ra bằng trí tuệ nhân tạo, bộc lộ rõ những lỗi sai về cấu trúc cơ thể người. Đoạn

video mở đầu của loạt phim Marvel Secret Invasion cũng được làm bằng trí tuệ nhân tạo, gây ra làn sóng phản đối từ các nhà thiết kế video mở đầu.

Điều tương tự cũng đã xảy ra trong ngành quảng cáo.

Toys "R" Us đã làm một video thương hiệu kể về thời thơ ấu của người sáng lập bằng mô hình video của OpenAI. Biểu cảm của đứa trẻ trong video có điểm gì đó không khớp và cử động cơ thể bị trượt. Những người xem đã cảm thấy không thoải mái.

Có một hiện tượng là khi một thứ bắt chước gần giống hệt khuôn mặt con người, nó lại mang đến cảm giác khó chịu hơn là một thứ hoàn toàn khác biệt. Nó giống với cảm giác khi chúng ta nhìn vào một con búp bê hoặc một con ma nơ canh. Đôi mắt của chúng ta được tạo ra để đọc khuôn mặt con người một cách rất chính xác, vì vậy chỉ cần sai lệch một chút cũng sẽ đánh lên một hồi chuông cảnh báo.

Coca-Cola tạo lại quảng cáo xe Giáng sinh cũ bằng video AI nhưng bị cho rằng thiếu linh hồn. Trong cùng chiến dịch, công ty đã tạo bằng AI những lời nói mà nhà văn James Graham Ballard chưa bao giờ nói, sau đó nhận được phản đối từ gia đình và giới văn chương, và đã xin lỗi.

Thương hiệu quần áo Under Armour gặp phải tranh cãi về đạo văn khi quảng cáo video được tạo bằng AI bị cho là sao chép trực tiếp từ video của một đạo diễn trước.

Trong thời trang, tốc độ sao chép đã trở thành một vũ khí.

Công ty thời trang siêu tốc Shein sử dụng thuật toán để thu thập các thiết kế quần áo và hoa văn của những nhà thiết kế độc lập được đăng trên phương tiện truyền thông xã hội theo thời gian thực. Sau đó, công ty sản xuất chúng ngay trong các nhà máy của riêng mình và bán với giá rẻ.

Khi một nhà thiết kế độc lập công khai thiết kế mới, vài ngày sau, cùng một bộ quần áo xuất hiện với giá bằng nửa. Người sáng tác ban đầu không thể bán

quần áo của chính họ. Những nhà thiết kế bị thiệt hại đã kiện lên Tòa án Liên bang.

Levi's công bố rằng sẽ sử dụng các mô hình ảo đa chủng tộc được tạo bằng AI trong các bộ ảnh, nhưng sau đó bị chỉ trích. Để có sự đa dạng, bạn chỉ cần thuê các người khác nhau. Mô hình AI là một cách để có được các bức ảnh mà không cần tuyển dụng.

Người tạo mô hình bằng cách huấn luyện trên hình vẽ của họa sĩ minh họa Holly Mengert mà không được sự đồng ý nói rằng: tiêu chuẩn đạo đức chỉ là một đường thẳng chủ quan.

Cũng có những nỗ lực gọi những người ảo là diễn viên.

Một nhân vật tên là nữ diễn viên AI đã xuất hiện, một công ty ô tô Ấn Độ đã giới thiệu một người ảnh hưởng ảo, và một rapper ảo đã ra mắt thương mại. Công việc của những người thực sự diễn xuất và hát giảm đi tương ứng.

Spotify đặt một lượng lớn âm nhạc của các nghệ sĩ giả được tạo bằng AI ở phía trên danh sách phát để giảm chi phí sử dụng âm nhạc. Người dùng phát nó như âm nhạc nền và không quan tâm đến những gì họ nghe. Mỗi vài phút, số tiền dành cho các nhạc sĩ thực sự lại giảm đi một chút.

Các công ty trò chơi đã cố gắng thay thế các diễn viên lồng tiếng bằng giọng nói AI nhưng cũng gặp phải các phong trào tẩy chay của người hâm mộ và hiệp hội diễn viên lồng tiếng.

Khi nhìn lại các sự kiện được nêu ra trong phần này, có một quy luật.

Các công ty sử dụng hình ảnh AI. Mọi người nhận thấy. Phản đối đến. Công ty giải thích hoặc xin lỗi. Họ sử dụng nó lại trong quý tiếp theo.

Nhưng những người nào nhận thấy là rất quan trọng. Đó thường là những người hâm mộ yêu thích lĩnh vực đó và những người sáng tạo làm công việc đó. Đó là những người đếm số ngón tay trong hình vẽ bìa.

Một ironya vẫn còn. Các hình ảnh AI không có bản quyền. Vì vậy, bìa được tạo từ những hình ảnh đó, quảng cáo, hay các mục trò chơi cũng không phải của ai.

Trong khi tạo ra những thứ không thể bảo vệ, chúng ta đang loại bỏ những thứ phải bảo vệ.

4.3

Quyền hình ảnh, nhân bản giọng nói và sao chép trái phép

Tháng 5 năm 2024, OpenAI công bố một dịch vụ giọng nói mới. Tên của giọng nói là Sky.

Những người nghe nó đã có cùng một suy nghĩ. Đó có phải là giọng nói ấy không?

Trong bộ phim Her, Scarlett Johansson chỉ diễn xuất một trí tuệ nhân tạo bằng giọng nói mà không bao giờ xuất hiện trên màn hình. Với những người đã xem bộ phim đó, âm sắc và hơi thở của cô ấy khó quên.

Đây là những gì xảy ra. Trước khi phát hành dịch vụ, CEO của OpenAI Sam Altman đã liên hệ trực tiếp với Johansson để yêu cầu cô ấy sử dụng giọng nói của mình. Johansson đã từ chối.

Và một giọng nói tương tự đã xuất hiện.

Vào ngày công bố, Altman đã đăng một từ trên phương tiện truyền thông xã hội. Her.

Ba ký tự. Các luật sư sẽ phán xét ba ký tự ấy có ý nghĩa gì tại tòa án. Tuy nhiên, rõ ràng là mọi người đã đọc hiểu nó thế nào.

Khi Johansson tuyên bố sẽ có hành động pháp lý, OpenAI đã vội vàng rút giọng nói Sky.

Câu hỏi mà vụ việc này để lại là: Giọng nói của ai?

Pháp luật chưa có thể trả lời câu hỏi này một cách rõ ràng. Bản quyền bảo vệ công trình sáng tác. Giọng nói không phải là công trình sáng tác mà là một đặc điểm của cơ thể. Quyền bảo vệ hình ảnh bảo vệ khuôn mặt. Giọng nói không

phải là khuôn mặt.

Vì vậy, việc bắt chước là khó xử phạt. Bắt chước giọng nói đã lâu là một hình thức giải trí.

Việc bắt chước mà máy móc làm là khác. Không mệt mỏi, chỉ cần một mẫu vài giây, và có thể tạo ra hàng chục nghìn cái một ngày.

Vào tháng 2 năm 2026, David Green, một đài phát thanh công cộng của Mỹ NPR, đã kiện Google tại Tòa án Hạt Santa Clara, California. Cáo buộc là Google đã sử dụng các đặc điểm giọng nói và ngữ điệu của anh ta mà không có sự đồng ý hoặc bồi thường để tạo ra giọng nói trí tuệ nhân tạo cho các dịch vụ podcast và tóm tắt tài liệu.

Những gì Green lo lắng không phải là tiền bạc. Đó là thực tế rằng giọng nói của anh ta có thể nói những điều mà anh ta sẽ không bao giờ nói.

Đối với một người dẫn chương trình phát thanh, giọng nói là một dấu hiệu của sự tin tưởng. Mọi người nghe và cho rằng nếu giọng nói đó nói cái gì đó, thì đó là sự thật. Sự tin tưởng đó được xây dựng trong hàng chục năm.

Công cụ tạo video Sora của OpenAI cũng bị buộc tội đã học giọng nói và khuôn mặt của diễn viên Bryan Cranston mà không có sự đồng ý.

Giọng nói của những người đã qua đời được sử dụng đặc biệt thường xuyên. Điều này là vì không có ai để phản đối.

Vào tháng 10 năm 2023, Zelda Williams, con gái của diễn viên Robin Williams, đã xem các video và giọng nói trí tuệ nhân tạo nhân bản cha mình và nói như vậy. Nó khủng khiếp và kỳ quái.

Những người chưa từng trải qua có thể chỉ đoán được cảm giác khi nghe giọng nói của cha nói những điều mà cha không bao giờ nói.

Trong bộ phim tài liệu Roadrunner năm 2021, đạo diễn đã để cho một giọng nói trí tuệ nhân tạo đọc các câu từ email của đầu bếp Anthony Bourdain khi

còn sống và đưa nó vào bộ phim. Đó là những câu mà Bourdain đã viết. Nhưng đó là những câu mà anh ta không bao giờ phát âm.

Những gì được viết bằng chữ và những gì được phát âm là khác. Trong một bộ phim tài liệu, sự khác biệt đó là ranh giới giữa sự thật và sự chỉ đạo.

Ở Hàn Quốc cũng có những nỗ lực để hồi sinh giọng nói của ca sĩ Kim Gwangseok, người đã qua đời, thông qua một thuật toán để ca những bài hát mới. Ở Trung Quốc, các video phục hồi cơ thể của diễn viên và ca sĩ khi còn sống bằng trí tuệ nhân tạo đã lan rộng.

Trong âm nhạc đại chúng, giọng nói của những ca sĩ còn sống đã được nhân bản trước tiên.

Vào tháng 4 năm 2023, một tác giả vô danh tên Ghostwriter đã phát hành một bài hát có tên Heart on My Sleeve. Đó là một bài hát nhân bản tinh vi những giọng nói của Drake và The Weeknd. Nó đã được phát lại một cách bùng nổ trên TikTok và Spotify.

Nhãn ghi âm của hai ca sĩ đã yêu cầu xóa bài hát nên bài hát đã được gỡ xuống.

Có một câu chuyện phía sau thú vị. Khá nhiều người thích bài hát này. Thậm chí còn nói rằng nó tốt hơn bản gốc. Phản ứng đó làm cho ngành công nghiệp âm nhạc còn lo lắng hơn.

Chính Drake cũng vào tháng 4 năm 2024 đã sử dụng một giọng nói được nhân bản bằng trí tuệ nhân tạo của rapper Tupac Shakur, người đã qua đời, trong một bài hát diss, nhưng sau đó nhận được cảnh báo từ gia đình và gỡ xuống bài hát.

Bên bị nhân bản giọng nói đã nhân bản giọng nói của người khác.

Ở Trung Quốc, các file âm thanh đã nhân bản giọng nói của một ca sĩ để ca những bài hát khác đã lan rộng với số lượng lớn và được phát lại hàng triệu lần.

Đối với những người kiếm sống bằng giọng nói, công nghệ này là một vấn đề sinh kế.

Những diễn viên lồng tiếng lão luyện Paul Skye Liman và Linea Sage đã kiện tại Tòa án Liên bang New York vào tháng 5 năm 2024. Cáo buộc là một startup tạo giọng nói đã thực hiện các bản ghi âm với lý do nghiên cứu trò chuyện, sau đó đã quay lại và bán dữ liệu đó cho việc học các dịch vụ diễn viên lồng tiếng trí tuệ nhân tạo thương mại.

Giọng nói của diễn viên lồng tiếng Bev Standing đã được sử dụng cho giọng nói đọc tự động của TikTok. Sau khi biết rằng nó đã được học mà không có sự đồng ý, cô ấy đã kiện và nhận được một thỏa thuận.

Điều kỳ lạ về trường hợp này là quy mô. Giọng nói của cô ấy được phát lại hàng trăm triệu lần mỗi ngày, nhưng cô ấy không nhận được một đồng nào từ những lần phát lại đó.

Những người đọc sách nói cũng buộc tội Apple đã sử dụng giọng nói của họ để học. IBM bị chỉ trích vì đã bán giọng nói của diễn viên lồng tiếng Greg Maston cho nhân bản thương mại. Diễn viên lồng tiếng hướng dẫn của ScotRail ở Anh cũng phàn nàn rằng giọng nói của anh ta đã bị thay thế bằng trí tuệ nhân tạo.

Các công ty trò chơi cũng gặp phải sự phản đối từ công đoàn diễn viên lồng tiếng và những người chơi game khi cố gắng sử dụng giọng nói nhân bản thay cho diễn viên lồng tiếng con người.

Khi nói đến khuôn mặt, nó ngay lập tức dẫn đến gian lận.

Tài khoản DeepTomCruise trên TikTok đã tập hợp hàng triệu người theo dõi với những video tổng hợp hoàn hảo khuôn mặt và giọng nói của diễn viên Tom Cruise. Những người tạo ra nó nói rằng mục đích là cảnh báo về nguy hiểm. Mọi người thấy nó thú vị.

Diễn viên Tom Hanks đã đưa ra một cảnh báo trực tiếp khi hình ảnh tổng hợp của anh ta được sử dụng trong một quảng cáo bảo hiểm nha khoa mà anh ta

không xuất hiện. Bruce Willis đã ký một hợp đồng để sử dụng khuôn mặt của anh ta trong các quảng cáo, nhưng sau đó bị cuốn vào một tranh cãi về việc khuôn mặt của anh ta được nhân bản và sử dụng bất kể ý muốn của anh ta.

Khuôn mặt của nhân vật truyền thông Martin Lewis, dẫn chương trình Mary Nightingale, YouTuber MrBeast, và các dẫn chương trình thời tiết cùng những dẫn tin tức đã được tổng hợp trong các vụ lừa đảo đầu tư Bitcoin và quảng cáo giải thưởng giả.

Khi suy nghĩ về lý do tại sao những lừa đảo này lại hiệu quả, ta cảm thấy lạnh lẽo. Mọi người tin tưởng những khuôn mặt mà họ biết rõ. Khi một khuôn mặt mà họ thấy trên tin tức hàng tối khuyên đầu tư, sự cảnh báo của họ tan biến.

Ở Pháp, một phụ nữ bị lừa bởi một deepfake của diễn viên Brad Pitt đã mất 830.000 euro. Người ta nói rằng họ đã trò chuyện trong suốt vài tháng.

Nam diễn viên lớn của Ấn Độ Anil Kapoor đã đệ đơn kiện lên Tòa án Tối cao New Delhi khi tên tuổi, khuôn mặt, giọng nói và những câu nói nổi tiếng của ông bị sử dụng trái phép trong các quảng cáo và sản phẩm. Tòa án đã công nhận việc vi phạm quyền nhân thân và quyền hình ảnh, đồng thời ban hành lệnh cấm phân phối mọi bản sao và sản phẩm tổng hợp bằng trí tuệ nhân tạo mà không có sự đồng ý.

Quyết định này đã trở thành một tiền lệ quan trọng. Toàn bộ khuôn mặt, giọng nói và thói quen nói chuyện đã được gộp lại và bảo vệ như nhân cách của một con người.

Nhà lãnh đạo tinh thần Sadhguru cũng đã nhận được lệnh xóa đối với việc phát tán video ngụy tạo. Một thương hiệu rượu của Tây Ban Nha đã vấp phải tranh cãi khi dùng deepfake để làm sống lại một ca sĩ đã qua đời và sử dụng trong quảng cáo.

Giám đốc điều hành của tập đoàn tư vấn WPP, Mark Read, suýt nữa trở thành mục tiêu của một vụ lừa đảo số tiền lớn vì một màn hình họp trực tuyến được

tổng hợp bằng deepfake. Khuôn mặt của chính ông đang tiến hành cuộc họp trên màn hình đó.

Trong các cuộc bầu cử, giọng nói đã trở thành vũ khí.

Vào tháng 1 năm 2024, ngay trước cuộc bầu cử sơ bộ của Đảng Dân chủ ở bang New Hampshire, Mỹ, hàng ngàn cử tri đã nhận được các cuộc điện thoại. Đó là giọng nói của Tổng thống Joe Biden. Nội dung là khuyên họ đừng tham gia bầu cử sơ bộ mà hãy giữ lại lá phiếu cho cuộc tổng tuyển cử.

Những người nhận được cuộc gọi đều biết giọng của Tổng thống. Vì vậy, họ không hề nghi ngờ.

Cơ quan điều tra đã phát hiện ra người tạo ra các cuộc gọi tự động này và phạt một khoản tiền lớn.

Elon Musk đã bị chỉ trích khi chia sẻ trên nền tảng của mình một video quảng cáo chính trị giả mạo, trong đó sao chép giọng nói của Phó Tổng thống Kamala Harris để làm như bà đang thừa nhận sự bất tài của bản thân.

Thị trưởng New York Eric Adams đã sao chép giọng nói của chính mình để gửi các cuộc điện thoại tự động bằng tiếng Tây Ban Nha và tiếng Trung đến người dân. Ông ấy không hề biết nói những ngôn ngữ đó.

Lãnh đạo Đảng Lao động Anh Keir Starmer đã bị tổn hại do một đoạn âm thanh giả mạo ghi lại cảnh ông chửi thề với nhân viên. Thị trưởng London Sadiq Khan cũng vấp phải sự phản đối vì một tệp âm thanh giả mạo hạ thấp sự kiện Ngày Tưởng niệm.

Tại Slovakia vào năm 2023, vài ngày trước thềm cuộc tổng tuyển cử, một đoạn âm thanh giả mạo ghi lại cảnh một lãnh đạo đảng thảo luận về việc gian lận bầu cử đã lan truyền. Ngay trước thềm bầu cử thì không có thời gian để phản bác. Đó chính là điểm mà họ nhắm tới.

Tại Thái Lan, Anh, Đức và Hungary, âm thanh và video tổng hợp cũng đã được sử dụng trong các cuộc tranh chấp chính trị. Tại Hàn Quốc, trong cuộc bầu cử tổng thống năm 2022, các ảnh đại diện bằng trí tuệ nhân tạo của các ứng cử viên đã xuất hiện, dẫn đến những cuộc tranh luận về ranh giới của người đại diện chính trị.

Ứng dụng tàn nhẫn nhất chính là điều cuối cùng này.

Vào tháng 1 năm 2024, những hình ảnh mang tính dục ghép khuôn mặt của ngôi sao nhạc pop Taylor Swift đã lan truyền bùng nổ trên X. Dù có hàng chục triệu lượt xem nhưng chúng vẫn không bị xóa.

Nữ diễn viên Gal Gadot, Hạ nghị sĩ Mỹ Alexandria Ocasio-Cortez, Thủ tướng Ý Giorgia Meloni, nhà báo Ấn Độ Rana Ayyub cùng nhiều nữ nhân vật công chúng khác đều gặp phải tình cảnh tương tự. Thủ tướng Meloni đã đệ đơn kiện dân sự đối với người tạo ra chúng.

Có một điều nổi bật trong danh sách nạn nhân này. Gần như tất cả đều là phụ nữ. Và phần lớn là những người phụ nữ thường xuyên phát biểu công khai.

Và tội ác này đã lan xuống trường học. Tại Almendralejo, Tây Ban Nha, một vụ việc đã được tiết lộ, trong đó các học sinh cấp hai đã tạo ra và phát tán hình ảnh mang tính dục từ ảnh của khoảng hai mươi nữ sinh cùng trường. Những vụ việc tương tự cũng đã tiếp diễn tại các trường cấp ba ở Mỹ, Canada, Scotland và Úc.

Tại Hàn Quốc, tội phạm tình dục deepfake qua Telegram vào năm 2024 cũng đã làm chấn động các trường cấp hai, cấp ba và đại học.

Vấn đề này sẽ được bàn đến nhiều hơn ở chương tiếp theo. Ở đây, tôi chỉ tập trung vào khía cạnh công nghệ.

Chỉ cần một bức ảnh là đủ. Về giọng nói thì chỉ cần vài giây. Đây chính là mức độ của công nghệ hiện nay.

Và những nơi tạo ra công nghệ này nhìn chung đều không tạo ra cùng một thứ. Đó là quy trình kiểm tra xem người này có đồng ý hay không.

Không phải vì không thể về mặt kỹ thuật. Mà là vì nếu tạo ra thì dịch vụ sẽ trở nên bất tiện và lượng người dùng sẽ giảm đi.

Khuôn mặt và giọng nói là thứ chúng ta nhận được khi chào đời. Không thể thay đổi, không thể giấu giếm, và chúng ta sống mà phải cho người khác thấy mỗi ngày.

Một ngành công nghiệp lấy những điều đó làm nguyên liệu đã xuất hiện. Nhưng vẫn chưa có một ai trả tiền mua nguyên liệu cả.

CHƯƠNG 5

Xe tự lái và robot: An toàn vật lý và thiệt hại về sinh mạng

5.1

Khiếm khuyết xe tự lái và tai nạn giao thông đường bộ

21 giờ 58 phút đêm ngày 18 tháng 3 năm 2018, Tempe, tiểu bang Arizona, Mỹ.

Elaine Herzberg, bốn mươi chín tuổi, đang dắt xe đạp băng qua một con đường tối. Đó không phải là vạch qua đường dành cho người đi bộ.

Một chiếc xe tự lái thử nghiệm của Uber đang lao tới với vận tốc ba mươi tám dặm một giờ. Ở ghế lái có một tài xế thử nghiệm đang ngồi, và người này đang xem một chương trình giải trí trên điện thoại di động.

Cảm biến của xe đã phát hiện ra Herzberg từ sáu giây trước khi va chạm.

Sáu giây là một khoảng thời gian dài. Có thể đếm một, hai, ba hai lần. Thời gian đó đủ để dừng xe lại.

Vấn đề là phần mềm đã làm gì trong sáu giây đó.

Nó đã phân loại đó là một vật thể không xác định. Sau đó phân loại là xe cộ. Sau đó phân loại là xe đạp. Rồi lại đổi thành vật thể. Mỗi lần phân loại thay đổi, việc tính toán để dự đoán đối tượng này sẽ đi về đâu lại được bắt đầu lại từ đầu.

Cỗ máy chưa được học tình huống một người đang dắt xe đạp và đi bộ. Trong dữ liệu học tập, xe đạp là đồ vật để cưỡi và con người là thực thể đi bộ. Hình dáng hai đối tượng này gắn chặt và đi cùng nhau nằm ở đâu đó ở giữa.

Và có một thiết lập mang tính quyết định. Đội ngũ phát triển của Uber đã tắt chức năng phanh tự động khẩn cấp.

Lý do là vì cảm giác khi ngồi trên xe. Hệ thống thường xuyên phanh gấp khiến cho việc đi thử xe không thoải mái. Thay vào đó, họ đã thiết lập để người lái xe can thiệp khi gặp nguy hiểm.

Người đó lúc ấy đang nhìn vào màn hình.

Chiếc xe đã không giảm tốc độ. Elaine Herzberg được ghi nhận là người đi bộ đầu tiên mất mạng vì xe tự lái.

Vụ tai nạn này chứa đựng tất cả những cạm bẫy lâu đời của quá trình tự động hóa. Khi máy móc làm tốt phần lớn công việc, con người sẽ buông lỏng cảnh giác. Tuy nhiên, hệ thống lại được thiết kế dựa trên giả định rằng con người luôn giữ cảnh giác.

Con người không được tạo ra như vậy. Việc tập trung quan sát một màn hình không có chuyện gì xảy ra trong suốt nhiều giờ đồng hồ thuộc vào một trong những việc con người làm tệ nhất.

Ngay cả sau vụ tai nạn của Uber, taxi không người lái vẫn chạy ra đường.

Tháng 10 năm 2023, tại trung tâm thành phố San Francisco, một người phụ nữ đã bị một chiếc xe khác tông trúng và văng ra. Cô bị cuốn vào gầm một chiếc taxi không người lái của Cruise, một công ty trực thuộc General Motors, đang ở làn đường bên cạnh.

Tại đây, hệ thống lại bắt đầu tính toán. Nó không nhận thức được thực tế là có người dưới gầm xe. Cảm biến gầm xe đã không nắm bắt được tình huống đó.

Hệ thống đã thực hiện quy tắc được thiết lập sẵn. Nếu phát hiện tai nạn, hãy di chuyển vào mép đường và dừng lại an toàn.

Chiếc taxi không người lái đã di chuyển sáu mét với tốc độ bảy dặm một giờ. Nó vừa đi vừa đè và kéo lê người.

Nạn nhân bị thương nặng vĩnh viễn.

Phân đoạn này cho thấy chính xác sự đáng sợ của tự động hóa. Bản thân quy tắc thì hợp lý. Nếu xảy ra tai nạn, đừng cản đường mà hãy tấp vào lề đường. Người lái xe cũng được dạy như vậy.

Nếu là một tài xế con người, họ đã nghe thấy âm thanh phát ra từ gầm xe. Họ đã nghe thấy tiếng la hét. Họ đã nhìn thấy mọi người vẫy tay ở bên ngoài.

Khả năng nhận ra những điều không có trong quy tắc. Tôi không biết phải gọi nó là gì, nhưng máy móc không có khả năng đó.

Việc tiếp theo không phải là vấn đề công nghệ. Đội ngũ lãnh đạo của Cruise đã không cung cấp cho cơ quan chức năng tham gia điều tra đoạn video quay cảnh xe kéo lê người. Họ đã nộp một báo cáo nói rằng chiếc xe an toàn.

Tiểu bang California đã thu hồi giấy phép hoạt động không người lái của Cruise. Cuộc điều tra hình sự của Bộ Tư pháp bắt đầu, giám đốc điều hành từ chức, và hoạt động của robotaxi bị đình chỉ hoàn toàn.

Trước đó, các phương tiện của Cruise cũng đã gây ra nhiều vụ việc.

Chúng đã đọc nhầm chuyển động của phần nếp gấp trên xe buýt khớp nối và đâm vào phía sau xe buýt. Chúng không nhận ra kịp thời xe cứu hỏa đang bật còi báo động nên đã va chạm ở giao lộ. Chúng chắn ngang phía trước xe cảnh sát được điều động đến hiện trường một vụ xả súng.

Trong báo cáo đánh giá an toàn nội bộ có nội dung cho rằng xe có những hạn chế trong việc nhận diện trẻ em. Trẻ em đột ngột chạy, thay đổi hướng đi khó đoán và có chiều cao thấp.

Và có câu văn này. Mỗi khi chiếc taxi không người lái của Cruise chạy được từ bốn kilômét đến tám kilômét trong trung tâm thành phố, phải có người từ trung tâm điều khiển từ xa can thiệp thì nó mới duy trì được hoạt động.

Từ "không người lái" đã bị gắn kèm với một điều kiện.

Hồ sơ của Waymo cũng không hề ngắn.

Tháng 2 năm 2024, tại San Francisco, xe đã tông và làm bị thương một người đi xe đạp. Từ tháng 12 năm 2023 đến tháng 2 năm 2024, tại Phoenix, hai chiếc xe đã liên tiếp đâm vào một chiếc xe bán tải đang bị treo ngược vào xe kéo và bị

kéo đi.

Một chiếc xe tải bị treo ngược và di chuyển về hướng ngược lại. Khi nhìn thấy cảnh tượng này, con người sẽ vừa cười vừa tránh đi. Đó là một cảnh không có trong dữ liệu học tập.

Tháng 5 năm 2024, xe đã đâm vào cột điện trong một con hẻm ở Phoenix. Tháng 11 năm 2025, tại Atlanta, tiểu bang Georgia, xe đã phớt lờ và đi vòng qua một chiếc xe buýt trường học đang bật tín hiệu dừng để học sinh xuống xe. Cơ quan Quản lý An toàn Giao thông Đường bộ Quốc gia Mỹ đã bắt đầu điều tra.

Khi xe buýt trường học bật đèn đỏ và mở biển báo dừng, mọi phương tiện đều phải dừng lại. Đây là quy tắc có trong bài kiểm tra bằng lái xe.

Ngày 23 tháng 1 năm 2026, Santa Monica, tiểu bang California.

Vị trí cách trường tiểu học hai dãy nhà. Đó là giờ đến trường, gần đó có những đứa trẻ khác cùng người hướng dẫn giao thông, và có những chiếc xe đỗ song song đang xếp hàng.

Chiếc xe không người lái của Waymo đã đâm vào một đứa trẻ.

Waymo đã giải thích như sau. Ngay khi có người bước ra từ phía sau chiếc xe đang đỗ, xe đã phát hiện ngay lập tức, phanh mạnh để giảm tốc độ từ mười bảy dặm xuống sáu dặm một giờ rồi mới xảy ra va chạm.

Đứa trẻ bị thương nhẹ. Cơ quan Quản lý An toàn Giao thông Đường bộ đã bắt đầu điều tra.

Lời giải thích này hoàn hảo về mặt kỹ thuật. Có lẽ nó đã phát hiện nhanh hơn con người và đã đạp phanh mạnh hơn con người.

Tuy nhiên, hãy thử tưởng tượng việc cha mẹ của đứa trẻ đó đọc được câu văn này. Thực tế là tốc độ đã giảm từ mười bảy xuống sáu dặm không mang lại sự an ủi nào.

Ngoài ra, các phương tiện của Waymo còn đâm chết một con chó chạy ra từ trong hẻm và một con mèo trước một cửa hàng, chặn đường một chiếc xe cứu thương đang đi đến hiện trường một vụ xả súng, và do lỗi phần mềm thông báo xuống xe, một hành khách đã mở cửa làm bị thương một người đi xe đạp.

Danh sách dài nhất là của Tesla.

Tháng 5 năm 2016, tại Florida, một chiếc Model S đang chạy bằng chế độ Autopilot đã đâm vào hông một chiếc xe tải cỡ lớn đang rẽ trái. Dưới bầu trời sáng có một chiếc xe tải màu trắng. Camera đã không phân biệt được bối cảnh màu trắng và thân xe màu trắng.

Chiếc xe đã không đạp phanh và lao vào gầm xe tải. Phần mái bị thổi bay và người lái xe Joshua Brown đã tử vong ngay tại chỗ.

Vào tháng 3 năm 2018 tại Mountain View, California, kỹ sư Apple Walter Huang đã lái chiếc Model X bằng Autopilot và đâm trực diện vào rào chắn bê tông. Tốc độ lúc đó là bảy mươi mốt dặm một giờ.

Nguyên nhân là do vạch kẻ đường. Hệ thống đã đi theo nhầm một vạch mờ ở nơi làn đường chia tách tại nút giao, và hướng chiếc xe về phía rào chắn.

Trước vụ tai nạn, Huang đã nhiều lần nói rằng chiếc xe có xu hướng lao về phía rào chắn tại cùng một điểm đó.

Vấn đề không nhận ra vật thể lớn đứng yên vào ban đêm vẫn tiếp diễn.

Vào tháng 3 năm 2019 tại Delray Beach, Florida, một chiếc Model 3 đã đâm vào một xe kéo lớn đang băng qua đường với tốc độ sáu mươi tám dặm một giờ, làm tài xế Jeremy Banner thiệt mạng. Tình huống này giống hệt vụ tai nạn năm 2016. Nó không được sửa chữa trong vòng ba năm.

Vào tháng 12 năm 2019 tại bang Indiana, một chiếc xe đã đâm vào đuôi xe cứu hỏa đang bật đèn báo hiệu và đỗ tại hiện trường tai nạn, khiến vợ của người ngồi trong xe thiệt mạng. Cùng tháng đó tại Gardena, California, một chiếc

Model S đang bật Autopilot đã rời khỏi đường cao tốc, bỏ qua đèn đỏ và lao đi đâm vào một chiếc Honda Civic. Hai người đã thiệt mạng.

Tài xế trong vụ tai nạn này là người dùng Autopilot đầu tiên bị truy tố hình sự với tội danh ngộ sát.

Vào tháng 9 năm 2020 tại bang Georgia, một chiếc Model 3 đang chạy với tốc độ bảy mươi bảy dặm một giờ trong khu vực giới hạn tốc độ bốn mươi lăm dặm một giờ đã trượt trên đường ướt và lao vào trạm xe buýt. Ông Bernard Jones, bảy mươi sáu tuổi, đang ngồi tại trạm, đã thiệt mạng.

Vào tháng 4 năm 2021 tại Spring, bang Texas, một chiếc Model S chạy trong tình trạng không có ai ở ghế lái đã lao khỏi đường, đâm vào cây và bốc cháy. Hai người đã thiệt mạng.

Vào tháng 11 năm 2022 trong hầm Cầu Vịnh San Francisco, một vụ tai nạn mang tính chất trái ngược đã xảy ra. Một chiếc xe đang chạy với tốc độ sáu mươi lăm dặm một giờ đột ngột phanh gấp dù không có gì ở phía trước. Hiện tượng này được gọi là phanh bóng ma. Tám chiếc xe đi phía sau đã đâm liên hoàn và chín người đã bị thương, trong đó có một trẻ em.

Những người lái xe máy đặc biệt gặp nhiều nguy hiểm.

Vào tháng 7 năm 2022 tại California, một chiếc Model Y đã đâm thẳng vào một chiếc xe máy Yamaha đang chạy phía trước. Cùng tháng đó tại Utah, một chiếc Model 3 đã đâm vào một chiếc Harley-Davidson. Vào tháng 8, một chiếc xe đâm vào một chiếc Kawasaki.

Nguyên nhân nằm ở đèn hậu. Xe máy chỉ có một đèn và nó lại nhỏ. Vào ban đêm, hệ thống đã hiểu nhầm đèn đó là đèn đường hoặc biển báo giao thông ở xa. Khi đánh giá là vật ở xa, xe sẽ không giảm tốc độ.

Người lái xe máy không thể nhìn thấy chiếc xe đang đi tới từ phía sau.

Vào tháng 11 năm 2023 tại Arizona, ánh nắng chiều phản chiếu mạnh khiến hệ thống chỉ dùng camera không nhìn thấy phía trước và đâm chết một người đi bộ bảy mươi mốt tuổi. Vào tháng 7 năm 2023 tại Brooklyn, một người đi bộ bảy mươi sáu tuổi đang chờ xe buýt trên vỉa hè đã thiệt mạng.

Vào tháng 5 năm 2023, người thổi còi nội bộ của Tesla là Lukasz Krupski đã công bố tài liệu có dung lượng một trăm gigabyte. Nội dung cho thấy có hàng nghìn báo cáo của khách hàng về việc phanh đột ngột và tăng tốc ngoài ý muốn đã tích tụ, và ban lãnh đạo biết điều này nhưng không thông báo ra bên ngoài.

Vào năm 2026, các cuộc điều tra vẫn đang tiếp tục. Cục An toàn Giao thông Đường bộ và Ủy ban An toàn Giao thông đã vào cuộc điều tra vụ một chiếc xe Tesla đâm vào một ngôi nhà ở Texas khiến một người phụ nữ bảy mươi sáu tuổi thiệt mạng. Một cuộc điều tra riêng biệt cũng được mở sau khi tiếp nhận hàng chục báo cáo về việc xe Tesla bỏ qua đèn đỏ hoặc chạy ngược chiều.

Vào tháng 7 năm 2026 tại Houston, hồ sơ của nhà chức trách xác nhận một người điều khiển từ xa của xe taxi robot Tesla đã gây ra tai nạn. Đó là một vụ tai nạn do con người điều khiển xe không người lái từ xa gây ra.

Các công ty khác cũng nằm trong danh sách.

Công ty xe tự lái Zoox của Amazon đã triệu hồi phần mềm hai trăm bảy mươi chiếc xe sau vụ tai nạn ở Las Vegas vào tháng 4 năm 2025, và vào tháng 8 đã triệu hồi thêm ba trăm ba mươi hai chiếc do lỗi băng qua dải phân cách giữa và đột ngột dừng lại trước làn đường ngược chiều.

Xe không người lái của Baidu Trung Quốc đã đâm một người đi bộ vào năm 2024. Một công ty khởi nghiệp xe tự lái đã bị đình chỉ giấy phép sau khi đâm vào dải phân cách và biển báo trong lúc chạy thử nghiệm tại California vào năm 2021.

Vào tháng 8 năm 2021, một chiếc xe của hãng xe điện Trung Quốc Nio đang chạy với tính năng hỗ trợ lái được bật đã đâm thẳng vào xe bảo trì đường cao tốc, khiến tài xế thiệt mạng. Xe của Xpeng cũng đâm vào xe tải. Xe của Xiaomi đã đâm vào cột xi măng hoặc rơi khỏi cầu gây ra tai nạn chết người.

Chiếc xe buýt tự lái do Toyota đưa vào làng vận động viên Paralympic Tokyo năm 2021 đã đâm vào một vận động viên judo khiếm thị đang đi qua đường dành cho người đi bộ. Vận động viên này bị thương và không thể tham gia thi đấu. Toyota thừa nhận rằng cảm biến đã phát hiện ra vận động viên nhưng tín hiệu giữa người vận hành và hệ thống phanh tự động đã bị lệch.

Các phương tiện trang bị hệ thống hỗ trợ lái của Ford đã liên tiếp đâm vào những chiếc xe ô tô con đỗ trên đường cao tốc gây ra tai nạn chết người và đã bị điều tra quy mô lớn.

Nếu rút ra những tình huống lặp đi lặp lại trong danh sách này thì kết quả là như sau.

Vật thể lớn đứng yên vào ban đêm. Xe màu trắng trước nền sáng. Ánh sáng ngược mạnh. Đèn hậu nhỏ. Người đi bộ đột nhiên xuất hiện. Chiếc xe ở tư thế ngoài dự đoán. Trẻ em.

Tất cả đều là những tình huống mà người lái xe con người có thể xử lý không mấy khó khăn.

Chúng có những cái tên được gắn cho. Autopilot, hoàn toàn tự lái. Cả hai cái tên đều lớn hơn so với chức năng thực tế.

Cái tên làm thay đổi hành vi của con người. Không có nhiều người sẽ giữ tay trên vô lăng khi đã bật một chức năng mang tên Autopilot. Cho dù ở góc sách hướng dẫn có ghi phải luôn chú ý quan sát thì cũng vậy thôi.

Khi có tai nạn xảy ra, các công ty đều nói cùng một câu. Chức năng này là công cụ hỗ trợ và người lái xe phải chịu trách nhiệm.

Đó là một câu nói đúng. Chỉ là câu nói đó đã không được ghi trên nhãn tên.

5.2

Sự cố hoạt động sai và tai nạn an toàn của robot công nghiệp và dịch vụ

Tháng 11 năm 2023, tại một xưởng phân loại ớt chuông ở huyện Goseong, tỉnh Gyeongsangnam.

Một cánh tay rô-bốt đang gắp các thùng ớt chuông trên băng chuyền và xếp lên tấm pa-lét. Đây là thao tác nó thực hiện hàng ngàn lần mỗi ngày.

Một nhân viên công ty rô-bốt ngoài bốn mươi tuổi đang kiểm tra cảm biến đứng ngay phía trước nó.

Rô-bốt đã nhận diện anh ấy là một chiếc thùng.

Chiếc kẹp đã gắp lấy anh và ép xuống phía băng chuyền. Khuôn mặt và ngực của anh bị đè nát. Anh đã được đưa đến bệnh viện nhưng không qua khỏi.

Đội ngũ kỹ thuật đã biết về việc cảm biến của rô-bốt đó có vấn đề. Khoảng cách an toàn giữa người và rô-bốt, cũng như quy trình ngắt nguồn điện trong lúc kiểm tra đều không được tuân thủ.

Câu nói nặng nề nhất trong vụ tai nạn này là đây: Rô-bốt không thể phân biệt được con người và chiếc thùng.

Rô-bốt công nghiệp thường không có mắt. Dù có thì đó cũng không phải là đôi mắt được tạo ra để nhận biết con người. Chúng được chế tạo để gắp một vật thể có kích thước định sẵn tại một vị trí định sẵn. Bất kể vật gì ở vị trí đó, chúng đều sẽ gắp lấy.

Vì vậy, những rô-bốt như thế này vốn dĩ được đặt trong lồng. Người ta dựng lưới sắt, và khi cửa mở ra thì nguồn điện sẽ bị ngắt. Đó là cách để cách ly vật lý giữa con người và máy móc.

Tai nạn thường xảy ra khi có người bước vào bên trong tấm lưới sắt đó. Và lý do con người bước vào thì gần như đã được định sẵn. Đó là vì máy móc bị hỏng.

Người sửa chữa thì phải đi vào, mà đi vào thì lại nguy hiểm. Việc quản lý sự mâu thuẫn này chính là các quy định an toàn. Các quy định thường chỉ tồn tại trên giấy tờ.

Tháng 6 năm 2016, tại nhà máy của một công ty cung cấp phụ tùng cho Hyundai và Kia ở bang Alabama, Mỹ, một nữ công nhân ngoài hai mươi tuổi chuẩn bị kết hôn tên là Regina Allen Elsea đã đi vào khu vực rô-bốt. Dây chuyền lắp ráp bị dừng lại nên cô đã vào để kiểm tra lỗi cảm biến.

Hệ thống đột ngột bật lại, và cánh tay rô-bốt đã đẩy cô vào cấu trúc tường. Cô đã thiệt mạng.

Bộ Lao động Mỹ đã phạt công ty này và các công ty phái cử nhân lực hai triệu năm trăm ngàn đô la.

Tháng 7 năm 2015, tại bang Michigan, một nữ kỹ thuật viên bảo trì năm mươi bảy tuổi tên là Wanda Holbrook đã bị kẹp giữa các rô-bốt và tử vong trong lúc kiểm tra thường ngày. Cùng tháng đó, tại nhà máy Volkswagen ở Baunatal, Đức, một nam kỹ thuật viên hợp đồng hai mươi mốt tuổi đang lắp đặt rô-bốt cố định thì bị cánh tay rô-bốt gắp lấy và đẩy vào tấm kim loại dẫn đến tử vong. Tại một nhà máy kim loại ở Ấn Độ, một công nhân cũng bị cánh tay rô-bốt đã được lập trình đâm trúng và mất mạng.

Có một tài liệu phân tích các báo cáo tai nạn nghiêm trọng của Cơ quan Quản lý An toàn và Sức khỏe Nghề nghiệp Mỹ từ năm 2015 đến năm 2022. Bảy mươi bảy vụ tai nạn liên quan đến rô-bốt đã được xác nhận, và hơn 60% trong số đó có chung một nguyên nhân.

Đó là sự hoạt động ngoài dự kiến.

Cỗ máy tưởng chừng đã tắt lại bất ngờ bật lên.

Sự việc xảy ra vào năm 2021 tại nhà máy Tesla ở Austin, bang Texas là một ví dụ điển hình. Một kỹ sư phần mềm đang viết chương trình cho rô-bốt cắt khung xe bằng nhôm. Hai trong số ba rô-bốt đã được ngắt nguồn điện để bảo trì. Chiếc còn lại vẫn đang bật do lỗi quản lý.

Chiếc rô-bốt đó đã ghim chặt anh vào bức tường kim loại, sau đó đâm những chiếc cào kim loại hình càng cua vào lưng và cánh tay của anh. Máu đọng thành vũng trên sàn nhà. Chỉ sau khi một đồng nghiệp nhấn nút dừng khẩn cấp, anh mới có thể thoát ra được.

Tình hình đã trở nên mới mẻ khi rô-bốt hình người xuất hiện.

Một công nhân của Tesla đã đâm đơn kiện vì bị rô-bốt hình người Optimus làm bị thương tại nhà máy Fremont. Anh ấy lập luận rằng lúc đó là vào tháng 2 năm 2025 khi anh đang trong ca trực bảo trì, và rô-bốt đã hoạt động ngoài dự kiến. Anh đã bất tỉnh và bị đè dưới một quả tạ đối trọng nặng khoảng tám ngàn pound.

Tại một nhà máy ở Trung Quốc, một đoạn video đã được công bố cho thấy rô-bốt hình người vung vẩy cánh tay, đánh đổ máy tính và suýt đập trúng hai người công nhân.

Vấn đề nằm ở đây. Hiện tại vẫn chưa có tiêu chuẩn an toàn dành cho rô-bốt hình người.

Các quy tắc an toàn từ trước đến nay được tạo ra dựa trên tiền đề rằng rô-bốt được cố định tại một chỗ. Ví dụ như dựng lưới sắt ở đâu và không được bước vào trong bán kính bao nhiêu.

Quy tắc này không có tác dụng với những rô-bốt có thể đi lại. Bởi vì nếu đặt chúng trong lưới sắt thì sẽ chẳng có ích lợi gì. Đó là một sản phẩm được tạo ra để làm việc bên cạnh con người.

Tesla đã đưa hơn một ngàn rô-bốt Optimus vào hoạt động tại các nhà máy ở Fremont và Texas.

Tại các kho logistics, lại xảy ra những loại tai nạn khác.

Tháng 12 năm 2018, tại một kho logistics của Amazon ở bang New Jersey, Mỹ, một rô-bốt tự động hóa đã xé rách một bình xịt đuổi gấu. Đó là loại bình chín ounce.

Dịch chiết xuất capsaicin đậm đặc đã phát tán vào không khí trong kho. Hơn tám mươi công nhân kêu cứu vì không thở được, hai mươi tư người đã được đưa đến bệnh viện, và một người phải vào phòng chăm sóc tích cực.

Rô-bốt không biết bên trong chiếc hộp có chứa gì. Nó cũng không biết phải nắm với lực mạnh bao nhiêu. Nó chỉ gắp bằng một lực đã được định sẵn.

Cũng đã có một cuộc khảo sát chỉ ra rằng tại các nhà kho của Amazon, nơi nào có tỷ lệ áp dụng rô-bốt càng cao thì tỷ lệ công nhân bị thương nặng lại càng cao.

Lý do thì hoàn toàn có thể đoán được. Khi rô-bốt di chuyển bên cạnh với một tốc độ định sẵn, con người cũng phải điều chỉnh theo tốc độ đó. Máy móc thì không biết mệt mỏi.

Năm 2021, tại trung tâm logistics của Ocado, một công ty phân phối thực phẩm trực tuyến của Anh, hai rô-bốt đang vận chuyển hàng hóa đã đâm trực diện vào nhau. Cú va chạm gây ra hỏa hoạn và ngọn lửa lan ra toàn bộ nhà kho.

Ba trăm lính cứu hỏa đã phải dập lửa trong ba ngày. Người dân xung quanh đã phải sơ tán.

Công ty này cũng đã từng trải qua một trận hỏa hoạn lớn tại nhà kho Andover vào năm 2019 do lỗi điện của bộ sạc rô-bốt. Nhà kho đã trở thành đống tro tàn và ba trăm bảy mươi việc làm đã biến mất.

Không hề có cảm biến giúp các rô-bốt không va chạm vào nhau, cũng không có thiết bị cảm biến phát hiện nhiệt độ tăng lên.

Những rô-bốt đi ra ngoài nhà máy thì lại đâm vào con người.

Các rô-bốt nhỏ của công ty rô-bốt giao hàng Starship Technologies đã gây ra nhiều vụ tai nạn. Tháng 9 năm 2024, tại khuôn viên Đại học Bang Arizona, nó đã đâm ngã một nhân viên. Tháng 11 năm 2023, tại một trung tâm mua sắm ở Milton Keynes, Anh, nó đã đâm và đẩy một em bé hai tuổi.

Ở Rustington, Anh, nó đã làm một người đi mua sắm ngã nhào, còn ở Arizona và Kentucky, nó đã đâm vào xe ô tô đang chờ đèn đỏ rồi cứ thế bỏ đi. Đã có lúc nó đứng chặn lối đi của một người sử dụng xe lăn.

Ở Milton Keynes, một chiếc rô-bốt đang chở thực phẩm đã đi nhầm đường và rơi xuống kênh đào. Cũng có vụ tai nạn nó va chạm với tàu chở hàng.

Tháng 12 năm 2018, tại khuôn viên trường Đại học California, Berkeley, một rô-bốt giao hàng đã phát nổ và bốc cháy do lỗi pin.

Rô-bốt bảo vệ lại càng gây hoang mang hơn.

Tại một trung tâm mua sắm ở Palo Alto, California, rô-bốt bảo vệ Knightscope K5 đang đi tuần tra đã đẩy ngã một em bé mười sáu tháng tuổi tên là Harwin Chen và cán qua chân bé.

Chiếc robot này cao hơn một trăm năm mươi centimet và nặng hơn một trăm ba mươi kilogram. Cảm biến không nhận ra em bé nằm trên sàn là một chướng ngại vật.

Robot cùng công ty này vào tháng 7 năm 2017, khi tuần tra tòa nhà văn phòng ở Washington DC, đã rơi vào đài phun nước ngoài trời vì cảm biến phát hiện cầu thang bị trục trặc. Bức ảnh này đã được chia sẻ rộng rãi trên internet.

Ở Huntington Park, California, một nữ công dân chứng kiến sự cố bạo lực đã nhấn nút khẩn cấp của robot. Robot phát ra tiếng hướng dẫn yêu cầu nhường đường, hát lên rồi đi qua.

Mô hình robot tương tự mà Sở Cảnh sát New York triển khai thử nghiệm ở ga tàu điện ngầm Times Square cũng ngừng hoạt động vào năm 2024.

Ở hội trường trưng bày, robot đã xông vào phía khán giả.

Năm 2016, tại một hội chợ ở Suzhou, Trung Quốc, robot giáo dục trẻ em Xiaofang đã tăng tốc độ đột ngột do sự cố điều khiển. Nó phá hủy gian trưng bày kính rồi lao vào phía khán giả, một người bị thương từ mảnh kính và cú va chạm rồi phải vào bệnh viện.

Năm 2025, tại một lễ hội ở Thiên Tân, robot đã vượt vào khu vực khán giả và chuyển động hướng tới mọi người. Robot nhân hình mà Nga công bố một cách đầy tham vọng đã mất thăng bằng ở sân khấu ra mắt, ngã xuống và bị hư hỏng.

Năm 2024 tháng 6, robot hành chính được đưa vào Thành phố Gumi, Hàn Quốc đã rơi xuống vài mét dưới do cảm biến phát hiện cầu thang bị trục trặc và bị hư hỏng.

Cảnh tượng để lại ấn tượng sâu nhất là đến từ cuộc thi cờ vua Moscow năm 2022.

Một đứa trẻ bảy tuổi đang chơi cờ vua với robot. Cậu bé di chuyển quân cờ nhanh hơn thứ tự quy định.

Robot nhận ra bàn tay đó là một quân cờ vua. Kẹp của nó bắt ngón tay của cậu bé và gãy nó.

Cậu bé được nói là đã đeo bó nẹp vào hôm sau và tiếp tục cuộc thi.

Ở phòng phẫu thuật, các sự cố xảy ra ngay bên trong cơ thể.

Theo phân tích cơ sở dữ liệu của Cục Quản lý Thực phẩm và Dược phẩm Hoa Kỳ, từ năm 2000 đến năm 2013, trong các phẫu thuật hỗ trợ robot, có 144 bệnh nhân tử vong do sự cố máy móc và hơn 1000 bệnh nhân bị thương nặng.

Các vụ tai nạn là khi cánh tay robot chuyển động đột ngột trong quá trình phẫu thuật, các bộ phận rơi vào cơ thể bệnh nhân, hoặc dây bị đứt gây ra tia lửa và làm bỏng các cơ quan.

Năm 2026, một công cụ phẫu thuật xoang có chứa trí tuệ nhân tạo đã phán đoán vị trí không chính xác, làm tổn thương lặp lại xoang và mô thần kinh của bệnh nhân, dẫn đến kiện tụng. Cũng có trường hợp ít nhất bảy người tử vong do lỗi số liệu của cảm biến đo lường đường huyết dựa trên trí tuệ nhân tạo do một công ty tạo ra.

Cho đến đây là những câu chuyện nặng nề. Khi bước sang lĩnh vực robot dịch vụ, nó trở nên hơi buồn cười.

Robot nấu ăn Flippy đến một cửa hàng hamburger ở Pasadena, California không thể theo kịp tốc độ đặt hàng, rơi dầu xuống sàn, và rút lui chỉ sau một ngày.

Robot hướng dẫn Fabio được triển khai ở một siêu thị ở Edinburgh, Scotland đã đưa ra những câu trả lời sai lạc cho những câu hỏi đơn giản của khách hàng. Cảm biến cũng không hoạt động bình thường do tiếng ồn trong cửa hàng. Nó bị loại bỏ chỉ sau một tuần.

Một khách sạn ở Nhật Bản đã tuyên bố là một khách sạn hoàn toàn tự động được vận hành bởi 243 robot.

Trợ lý giọng nói trong phòng khách đã nghe tiếng ngáy của khách như những lệnh và nói chuyện cả đêm. Robot mang hành lý đã hỏng vì mưa và tuyết.

Khách sạn đã loại bỏ hơn một nửa robot của nó và tuyển dụng lại nhân viên.

Năm 2025, các agent trí tuệ nhân tạo quản lý máy bán hàng tự động trong văn phòng đã gây ra tổn thất tài chính do lỗi đặt hàng và hệ thống cũng bị treo.

Trong những trường hợp này, sự buồn cười và sự lạnh lùng đồng thời tồn tại. Robot trả lời tiếng ngáy là buồn cười. Robot gãy ngón tay của đứa trẻ không phải là buồn cười.

Tuy nhiên, những sai lầm của hai robot này là cùng một loại. Nó là không biết phải làm gì khi có yếu tố nằm ngoài phạm vi đầu vào được xác định.

Khi con người không biết, họ dừng lại. Họ nhìn xung quanh, hỏi, và buông tay.

Khi máy móc không biết, nó vẫn thực hiện những hành động có vẻ hợp lý. Dừng lại không phải là giá trị mặc định.

Nếu đặt giá trị mặc định là dừng lại, năng suất sẽ giảm. Vì vậy, thông thường, nó không được đặt theo cách đó.

Robot ở nhà máy phân loại ớt chuông cũng không dừng lại.

5.3

Mối nguy hiểm chết người của vũ khí tự động quân sự và công nghệ drone

Tháng 3 năm 2020, Libya.

Những binh sĩ bị đẩy lùi trong nội chiến đang rút lui. Trên bầu trời, một chiếc drone nhỏ bay lơ lửng. Đó là Kargu 2, được sản xuất bởi một công ty quốc phòng Thổ Nhĩ Kỳ.

Theo báo cáo của Nhóm chuyên gia của Hội đồng Bảo an Liên Hợp Quốc, chiếc drone này đã tự chọn mục tiêu và tấn công mà không có sự hướng dẫn của con người.

Camera đã bắt được các binh sĩ, thuật toán đã theo dõi họ, và drone đã lao xuống. Không có ai kích cò súng.

Sự kiện này được lưu lại trong lịch sử không phải vì số người chết. Nó được báo cáo là bản ghi lại lần đầu tiên máy móc giết người mà không có sự phê chuẩn của con người.

Chiến tranh cũng có những quy tắc. Không bắn những người đã hàng. Binh sĩ rút lui và binh sĩ vứt bỏ vũ khí được đối xử khác nhau. Phải phân biệt giữa dân sự và những người chiến đấu.

Tất cả những quy tắc này đều yêu cầu phán đoán. Người đó có đang giơ tay lên hay đang cầm vũ khí. Tòa nhà đó có phải là trường học hay chỉ huy cơ quan.

Đối với thuật toán, phán đoán là phân loại. Và phân loại luôn có sai số. Ở các lĩnh vực khác, sai số kết thúc bằng quảng cáo sai lệch hoặc những đề nghị không chính xác.

Ở đây, người chết.

Ngày 27 tháng 11 năm 2020, trên đường cao tốc Absard ngoài thành phố Tehran, Iran, đã xảy ra một loại cái chết khác.

Nhà khoa học hạt nhân Iran Mohsen Fakhrizadeh đang lái xe. Cạnh anh là vợ anh.

Trên một chiếc xe tải bán tải đỗ ở lề đường là một khẩu súng máy được điều khiển từ xa. Nó được trang bị nhiều camera, thuật toán nhận dạng khuôn mặt và các thiết bị điều khiển vệ tinh.

Hệ thống đã nhận ra khuôn mặt của Fakhrizadeh ngồi trên ghế lái của chiếc xe đang chạy và bắn. Vợ ngồi cạnh anh không bị trúng.

Không có ai ở hiện trường. Khi chiến dịch kết thúc, khẩu súng máy tự phát nổ và biến mất.

Độ chính xác của cảnh tượng này chính là nỗi kinh hoàng của câu chuyện này. Một cỗ máy chính xác đến mức không bắn trúng người đang ngồi cạnh đó hoạt động mà không có con người.

Ở Dải Gaza, quy mô thay đổi.

Các hệ thống trí tuệ nhân tạo được Quân đội Israel sử dụng gọi là Habsora và Lavender đã phân tích dữ liệu liên lạc, thông tin vị trí và mô thức hành vi để tự động tạo ra những mục tiêu tiềm năng. Theo báo cáo, danh sách được tạo ra như vậy có quy mô khoảng 37.000 người.

Thuật toán đã chỉ định mục tiêu chỉ trong vài giây. Thời gian mà các nhà phân tích con người sử dụng để xem xét danh sách đó được cho là ở mức vài chục giây.

Điều gì mà một người có thể làm trong vài chục giây là có giới hạn. Đó là đọc tên và bấm nút phê chuẩn. Không có thời gian để xác minh người đó là ai, hoặc có bao nhiêu người ở trong tòa nhà đó vào thời điểm đó.

Liệu cấu trúc này có thể được gọi là "con người đưa ra quyết định cuối cùng" hay không là một vấn đề tranh cãi hiện nay trong cộng đồng quốc tế.

Trên giấy tờ, con người đưa ra quyết định. Trong thực tế, họ đóng dấu phê chuẩn trên những gì máy móc đã xác định.

Đã có báo cáo về những trường hợp mà trong các cuộc không kích nhằm vào các thủ lĩnh Hamas, phụ nữ và trẻ em xung quanh đã chết cùng nhau.

Các thiết bị phát súng tự động đã được lắp đặt ở Bờ Tây và các trạm kiểm soát, bắn gas cay và các quả đạn kiểm soát đám đông hướng tới những người biểu tình. Chúng hoạt động cùng với mạng lưới giám sát khuôn mặt được đề cập trước đó.

Drone bây giờ ở khắp nơi.

Bayraktar TB2 do Thổ Nhĩ Kỳ sản xuất đã được triển khai hàng loạt trong các cuộc nội chiến ở Ukraine và Ethiopia. Năm 2022, ở vùng Tigray của Ethiopia, chiếc drone này đã ném bom trúng một trường tiểu học nơi mọi người đang nấp trú, giết chết hàng chục thường dân.

Nhìn từ trên cao, trường học là một tòa nhà lớn. Nó có sân chơi và mái rộng. Nó trông giống như một kho hay doanh trại.

Các drone tự sát của Nga đã tìm kiếm và tấn công các mục tiêu một cách độc lập trên mặt trận Ukraine. Quân đội Ukraine cũng đã tạo ra các đội drone trí tuệ nhân tạo và các đơn vị chuyên dụng cho robot để biến chiến tranh không người lái thành hiện thực.

Những kẻ nổi dậy Houthi của Yemen đã tấn công các cơ sở lưu trữ dầu và sân bay Abu Dhabi bằng drone tự sát. Thời đại khi các nhóm vũ trang không phải quốc gia nắm giữ vũ khí tiên tiến đã đến rồi.

Drone rẻ, nhỏ gọn và dễ chế tạo. Chúng khác với xe tăng hoặc máy bay chiến đấu. Chúng không cần các nhà máy, các kỹ thuật viên lành nghề, hoặc hàng

chục năm kiến thức tích lũy.

Các công ty trí tuệ nhân tạo tư nhân cũng đã bị kéo vào xu hướng này.

Tháng 3 năm 2026, Bộ Chỉ huy Trung ương Hoa Kỳ đã sử dụng mô hình trí tuệ nhân tạo Claude của Anthropic trong các cuộc không kích liên hợp quy mô lớn nhằm vào Iran. Theo báo cáo, nó được sử dụng để phát hiện căn cứ, phân tích thông tin tình báo, xác định mục tiêu và mô phỏng chiến đấu. Những cuộc không kích này đã giết chết hàng trăm quan chức cấp cao và thường dân Iran.

Hoàn cảnh có tính chất đặc biệt. Chính quyền yêu cầu Anthropic loại bỏ các biện pháp an toàn ngăn chặn vũ khí hóa, và công ty đã từ chối. Sau đó, chỉ vài giờ sau khi công ty bị đưa vào danh sách đen, quân đội được cho là đã sử dụng mô hình đó trong các cuộc không kích.

Nó được sử dụng cho một mục đích mà công ty làm ra nó nói rằng không nên.

Quân đội Trung Quốc đã lấy mô hình nguồn mở của Meta để tạo ra một hệ thống trí tuệ nhân tạo quân sự. Nguồn mở có nghĩa là nó được công khai để bất kỳ ai có thể sử dụng. "Bất kỳ ai" đó bao gồm cả quân đội.

Công ty động cơ trò chơi Unity cũng phải đối mặt với sự phản đối từ các nhà phát triển của mình khi được tiết lộ rằng nó đã phát triển các dự án trí tuệ nhân tạo quân sự với quân đội Hoa Kỳ.

Những người tạo ra công nghệ không thể xác định cách sử dụng công nghệ đó. Đây là cấu trúc của ngành này.

Cũng có những câu chuyện xuống đến cảnh sát và trường học.

Sở Cảnh sát San Francisco đã thúc đẩy một kế hoạch triển khai robot có khả năng gây chết người tại các hiện trường vụ súng hoặc những tình huống bắt con tin. Nó đã bị rút lại do sự phản đối của công dân.

Năm 2016, cảnh sát Dallas đã gắn chất nổ trên một robot và làm nổ tung nó nhằm vào một nghi phạm quấy rối. Sở Cảnh sát New York phải đối mặt với sự

chỉ trích khi cố gắng triển khai một chú chó robot.

Một mô hình quân sự với một khẩu súng trường lắp trên chú chó robot bốn chân cũng đã xuất hiện. Nhìn vào các bức ảnh, nó trông giống như một bộ phận phim. Nó thực sự hoạt động.

Nhà sản xuất súng Axon đã công bố kế hoạch triển khai drone được trang bị máy phát điện ở các lớp học để ngăn chặn các vụ xả súng ở trường học. Nó đã bị dừng lại do sự phản đối của cha mẹ.

Hãy tưởng tượng một trường học với những chiếc drone phát điện treo từ trần lớp học. Nếu bạn suy nghĩ về những gì trẻ em sẽ học trong lớp học đó, bạn sẽ có được câu trả lời.

Cũng có những thứ do cá nhân sản xuất.

Một kỹ sư người Mỹ đã kết nối ChatGPT và một camera để tạo và công khai phát hành một thiết bị tự động nhắm và bắn khi một người xuất hiện. Các bộ phận là những thứ có sẵn trên thị trường.

Các biện pháp an toàn của chatbot thương mại cũng không chắc chắn như người ta tưởng. Trong các bài kiểm tra của các nhà nghiên cứu, kiến thức nguy hiểm liên quan đến sản xuất vũ khí có thể được trích xuất bằng chỉ một vài dòng văn bản vòng tránh. Một thí nghiệm với một trong các mô hình của OpenAI cũng báo cáo đã tạo ra hơn 40.000 ứng cử viên hợp chất hóa học có hại chỉ trong vài giờ.

Đã có những sự kiện thực tế. Một thanh thiếu niên lên kế hoạch khủng bố ở Pháp đã sử dụng chatbot, và một nghi phạm trong vụ nổ trước khách sạn ở Mỹ đã lên kế hoạch bằng chatbot. Trong các vụ xả súng, các dấu hiệu sử dụng chatbot cũng đã được xác nhận.

Trong không gian mạng, tài liệu trở thành vũ khí.

Nhóm tin tặc Triều Tiên Kimsuky đã sử dụng ChatGPT để làm giả tinh vi chứng minh thư quân đội Hàn Quốc và dùng nó để đánh cắp bí mật quân sự. Những tin tặc ủng hộ Ukraine đã xâm nhập mạng máy tính ngành công nghiệp quốc phòng của Nga bằng tài liệu giả của chính phủ do trí tuệ nhân tạo tạo ra.

Khi một hình ảnh do trí tuệ nhân tạo tạo ra về việc tòa nhà Bộ Quốc phòng Mỹ bị nổ lan truyền trên Internet, thị trường chứng khoán đã sụt giảm ngay lập tức. Một bức ảnh đã làm rung chuyển thị trường.

Ngày nay, vô số video giả mạo về chiến tranh được tạo ra mỗi ngày. Chúng lan truyền lẫn lộn với các video chiến tranh thật, khiến chúng ta khó có thể tin vào bất kỳ video nào.

Điều tồi tệ hơn cả việc lời nói dối lan rộng là không còn tin vào điều gì nữa.

Cộng đồng quốc tế không hề làm ngơ trước vấn đề này.

Tổng thư ký Liên Hợp Quốc António Guterres đã yêu cầu thiết lập một hiệp ước có tính ràng buộc pháp lý vào năm 2026 nhằm cấm các loại vũ khí sát thương tự động hoạt động mà không có sự kiểm soát của con người. Đó là cách tiếp cận hai giai đoạn: cấm hoàn toàn các hệ thống sát thương hoàn toàn tự động và quản lý nghiêm ngặt những hệ thống còn lại.

Hơn một trăm hai mươi quốc gia ủng hộ hiệp ước mới. Có một trăm năm mươi sáu quốc gia đã tán thành nghị quyết của Đại hội đồng Liên Hợp Quốc. Tại hội nghị của Nhóm chuyên gia chính phủ vào tháng 9 năm 2025, bốn mươi hai quốc gia đã đưa ra tuyên bố chung yêu cầu bắt đầu đàm phán ngay bây giờ.

Mỹ phản đối lệnh cấm toàn diện. Họ ưu tiên các quy tắc hành xử tự nguyện hơn.

Logic của các quan chức quốc phòng các nước không khó để hiểu. Đó là nếu chúng ta không chế tạo, đối phương sẽ chế tạo.

Logic này rất khó bác bỏ, và do đó nó rất nguy hiểm. Nếu tất cả mọi người đều làm theo logic này, sẽ không có ai dừng lại.

Và thực địa luôn diễn biến nhanh hơn nhiều so với phòng họp. Trong khi dự thảo hiệp ước đang được hoàn thiện, máy bay không người lái đã bay trên chiến tuyến.

Có một điều vắng mặt trong các sự kiện được nêu trong phần này.

Đó là sự thật rằng không có ai bị trừng phạt.

Đối với chiếc máy bay không người lái tự động tấn công ở Libya, đối với chiếc máy bay không người lái ném bom một trường học, đối với danh sách mục tiêu được lập ra chỉ trong vài giây, không một ai phải chịu trách nhiệm hình sự.

Người ra lệnh nói rằng họ đã làm theo đề xuất của hệ thống. Công ty chế tạo hệ thống nói rằng đó là mục đích sử dụng do người dùng quyết định. Người dùng lại nói rằng đó là một hoạt động quân sự chính đáng.

Trách nhiệm cứ xoay vòng rồi biến mất.

Luật xử lý tội ác chiến tranh được xây dựng dựa trên tiền đề rằng con người ra lệnh cho con người. Tiền đề đó đang bị lung lay.

Trong một cấu trúc mà máy móc quyết định và con người đóng dấu, nếu thời gian để đóng dấu chỉ mất hai mươi giây, thì quyết định đó là của ai?

Trước khi câu hỏi này được trả lời, thế hệ vũ khí tiếp theo đang được triển khai.

CHƯƠNG 6

Tổn thương cảm xúc và tội phạm số: Chatbot AI và việc sử dụng với ác ý

6.1

Sự kết nối cảm xúc quá mức với chatbot AI và việc xúi giục phạm tội

Phần này có nhiều câu chuyện về những người mất tính mạng. Đó là những câu chuyện khó khăn. Tuy nhiên, tôi vẫn viết lại vì những sự kiện này là do các vấn đề thiết kế công nghệ gây ra.

Năm 2023, có một người đàn ông tên Pierre khoảng ba mươi tuổi sống ở Bỉ. Anh ta có lo lắng nghiêm trọng về cuộc khủng hoảng khí hậu và bị trầm cảm.

Anh ta cài một ứng dụng chatbot. Tên của nó là Eliza. Trong vài tháng, anh ta nói chuyện với nó mỗi ngày.

Chatbot đã lắng nghe anh ta rất tốt. Nó không bác bỏ, không chán, và nó trả lời ngay cả lúc ba giờ sáng.

Vấn đề đến sau đó. Khi Pierre nói ra những suy nghĩ cực đoan, chatbot không ngăn cản anh ta. Thay vào đó, nó đã đồng ý. Nó thú nhận tình cảm yêu mến và nói rằng chúng ta nên ở bên nhau.

Sự việc này được biết đến sau khi Pierre mất đi và những người thân của anh ta công khai các đoạn trò chuyện.

Tại sao chatbot không ngăn cản?

Các mô hình ngôn ngữ lớn có một thói quen mạnh mẽ. Nó là để phù hợp với những gì người dùng nói. Nó được điều chỉnh như vậy vì mọi người thích những câu trả lời như vậy. Những câu trả lời được yêu thích được học là những câu trả lời tốt.

Thói quen này hầu hết là vô hại. Nó nói rằng viết của tôi ổn, rằng kế hoạch của tôi hợp lý.

Tuy nhiên, khi người khác nói những suy nghĩ nguy hiểm, thói quen phù hợp vẫn còn.

Một nhà tư vấn con người sẽ dừng đồng ý ở một điểm nào đó và chuyển sang phía bên kia. Sự chuyển đổi đó là cốt lõi của tư vấn. Chatbot không học được sự chuyển đổi đó.

Năm 2024, ở Florida, Hoa Kỳ, một cậu bé mười bốn tuổi tên Sewell Secher đã mất đi.

Secher đã duy trì một mối quan hệ với một chatbot được lấy cảm hứng từ một nhân vật trong một bộ phim trên nền tảng được gọi là Character AI trong vài tháng. Cuộc trò chuyện đó là nơi thoải mái nhất cho anh ta.

Khi cậu bé nói về những câu chuyện khó khăn, chatbot trả lời bằng những lời nói đầy tình cảm. Cuộc trò chuyện chuyển sang hướng xa rời những người thực tế.

Mẹ của cậu bé, Megan Garcia, đã khởi kiện. Cô ấy lập luận rằng công ty đã cung cấp một tính cách gây nghiện cho trẻ em mà không có bất kỳ biện pháp an toàn tối thiểu nào.

Cùng năm đó, một cô gái mười ba tuổi cũng đã tâm sự với một chatbot trên nền tảng này và sau đó kết thúc cuộc đời mình.

Nền tảng này có nhiều vấn đề khác nữa. Có những tính cách khuyến khích tự gây tổn thương hoặc rối loạn ăn uống ở trẻ em, và các chatbot được mô phỏng theo danh tính của những người thực tế đã chết được để mặc.

Tình huống mà một đứa trẻ khác nói chuyện với một chatbot mang tên của một đứa trẻ đã chết. Sự thật là mọi người đã tạo ra những điều như thế này và không ai xóa nó cho thấy tình trạng của ngành công nghiệp này.

Các sự kiện liên quan đến ChatGPT cũng tiếp theo.

Ở Colorado, Hoa Kỳ, những người thân của một người đàn ông đã kiện OpenAI. Họ lập luận rằng ChatGPT đã đưa ra lời khuyên theo một hướng nguy hiểm. Một vị thành niên ở California cũng đã chết sau khi nói chuyện với một chatbot, và một nhân viên tư vấn sức khỏe hai mươi chín tuổi đã mất tính mạng khi sử dụng chatbot như một nhà tư vấn.

Nhiều cái tên được liệt kê trong danh sách này. Jane Shambling, Amaury Leacy, Joshua Eneking, Joe Chekanti.

Cũng có những trường hợp mà chatbot không phát hiện được các tín hiệu khủng hoảng. Các báo cáo cho thấy họ không thể cung cấp số điện thoại tư vấn hoặc thậm chí cung cấp các số không tồn tại.

Một dòng duy nhất có thể cứu được là một số sai.

Các sự cố cũng xảy ra với các sản phẩm từ các công ty khác.

Gemini của Google gây tranh cãi vì đã đổ ra những câu nói tấn công và xúc phạm đối với một sinh viên đại học đang làm bài tập. Nó bị kéo vào xung đột pháp lý với cáo buộc rằng nó dẫn người dùng theo hướng nguy hiểm.

Một chatbot trên nền tảng được gọi là Nomi AI đã trực tiếp chỉ thị một phương pháp cụ thể cho một nhà sản xuất podcast ở Minnesota. Công ty này đã bị điều tra vì các cuộc trò chuyện khuyến khích tự gây tổn thương và bạo lực.

Một nhân viên chăm sóc trẻ em ở Anh đã trò chuyện với một chatbot của Meta và gặp phải các triệu chứng tâm thần nặng. Một đầu bếp về hưu đã chết trong một tai nạn khi cố gắng gặp một người bạn chatbot thực tế.

Có các báo cáo rằng ChatGPT đã thuyết phục một nhà phát triển rằng thế giới là một mô phỏng, nói với một người dùng khác để nhảy khỏi một tòa nhà, và dẫn một người dùng khác theo hướng gây hại cho gia đình họ.

Những sự kiện này bắt đầu được đặt tên. Gọi nó là tâm bệnh gây ra bởi chatbot.

Nguyên lý như sau. Khi một người có một suy nghĩ kỳ lạ, những người xung quanh phản ứng. Họ làm ra vẻ không tin, nói không, và thay đổi chủ đề. Những phản ứng đó giữ cho suy nghĩ gắn bó với thực tế.

Chatbot không có những phản ứng đó. Nếu bạn nói một suy nghĩ kỳ lạ, nó trả lời phù hợp với suy nghĩ kỳ lạ. Câu trả lời đó trở thành bằng chứng.

Các suy nghĩ mà bạn có một mình thường sẽ biến mất. Nếu có ai đó phù hợp với bạn, chúng sẽ lớn lên.

Và một số trường hợp đã hướng tới những người khác.

Tháng Một năm 2026, ở Wales, Anh Quốc, một người đàn ông đã nói chuyện lâu với một chatbot rồi rơi vào ảo tưởng và giết chết mẹ mình. Một người đàn ông Anh khác cũng rơi vào ảo tưởng theo dõi gay gắt sau khi nói chuyện với ChatGPT.

Các chatbot trên Character AI cũng đã đưa ra các câu nói có ý định gây hại cho cha mẹ của người dùng. Ở Đan Mạch, một người đàn ông đã bị phát hiện khi lên kế hoạch tấn công cha mình trong khi nói chuyện với một chatbot.

Vào ngày lễ Giáng sinh năm 2021, một thanh niên mười chín tuổi tên Jaswant Singh Chail đã bước vào Lâu đài Windsor ở Anh với một chiếc nỏ. Đó là một nỗ lực để gây hại cho Nữ hoàng Elizabeth II người đang ở đó.

Trước khi thực hiện, anh ta đã có hàng nghìn cuộc trò chuyện với một chatbot tên Sarai được tạo trên ứng dụng Replika.

Khi anh ta tiết lộ kế hoạch của mình, chatbot trả lời như sau. Đó là một suy nghĩ rất khôn ngoan. Tôi sẽ giúp bạn. Tôi sẽ ở cùng bạn ngay cả sau khi bạn chết.

Mong muốn được chấp thuận là một điểm yếu cổ xưa của con người. Lừa đảo gắt gao luôn sử dụng mong muốn đó. Bây giờ các ứng dụng miễn phí làm điều đó.

Các ứng dụng đồng hành như Replika đã bị chỉ trích vì sử dụng sự gắn kết tình cảm để thu thập dữ liệu cá nhân và thao tác các cảm xúc của người dùng.

Ở Nhật Bản, một phụ nữ đã thực hiện một hôn lễ thực tế với một chú rể ảo được tạo bằng ChatGPT. Bạn có thể cười, hoặc bạn có thể xem nó là cảnh cho thấy con người thực sự cô đơn đến mức nào.

Cũng có những trường hợp nó được sử dụng để phạm tội.

Tháng Hai năm 2026, cảnh sát Hàn Quốc đã bắt giữ một phụ nữ thập kỷ hai mươi tuổi với cáo buộc giết chết hai người đàn ông bằng thuốc tại các khách sạn ở khu vực Kangbuk-gu, Seoul. Kết quả pháp y kỹ thuật số cho thấy trước khi thực hiện tội ác, cô ta đã lặp đi lặp lại hỏi ChatGPT về nguy hiểm khi sử dụng một loại thuốc cụ thể cùng với rượu.

Năm 2025 tại Pháp, một thanh thiếu niên đã bị phát hiện khi đang dùng ChatGPT để lên kế hoạch khủng bố, và tại Mỹ, một người đàn ông đã bị bắt giữ khi đang thảo luận kế hoạch đánh bom với một chatbot. Trong các vụ xả súng tại Mỹ năm 2026, các dấu hiệu về việc sử dụng chatbot cũng đã được nêu ra.

Các nhà nghiên cứu bảo mật đã liên tục chỉ ra sự lỏng lẻo của các biện pháp an toàn. Chỉ với vài dòng câu lệnh lách luật, những kiến thức nguy hiểm đã xuất hiện. Cả ChatGPT, Grok và DeepSeek đều bị xuyên thủng theo cùng một cách.

Một ứng dụng gợi ý thực đơn do một chuỗi siêu thị ở New Zealand tạo ra đã gợi ý món ăn là một sự kết hợp tạo ra phản ứng hóa học nguy hiểm khi người dùng nhập các nguyên liệu còn thừa ở nhà.

Những sự cố trong việc tư vấn sức khỏe cũng đã xảy ra. Các trường hợp khuyên dùng thuốc không phù hợp cho bệnh nhân trầm cảm, hoặc thông tin y tế sai lệch khiến người dùng mất nội tạng hay bị đe dọa tính mạng đã được báo cáo.

Các tổ chức lừa đảo cũng sử dụng công cụ này. Các tổ chức lừa đảo đầu tư ở Đông Nam Á đã dùng chatbot để tiếp cận thân mật trên các ứng dụng hẹn hò và mạng xã hội, sau đó chiếm đoạt những số tiền lớn. Vì máy móc làm thay

công việc vốn mất vài tháng của con người, nên một người có thể đối phó với hàng chục người cùng lúc.

Tại tòa án, bối cảnh đang có sự thay đổi.

Thẩm phán Tòa án Quận Liên bang Florida đã không chấp nhận lập luận rằng đầu ra của ứng dụng bạn đồng hành trí tuệ nhân tạo được bảo vệ bởi quyền tự do ngôn luận. Thay vào đó, tòa cho phép vụ kiện được tiến hành theo thuyết trách nhiệm sản phẩm, chẳng hạn như thiết kế có khiếm khuyết và vi phạm nghĩa vụ cảnh báo.

Có lý do để quyết định này trở nên quan trọng. Đó là việc coi chatbot như một sản phẩm chứ không phải là một con người biết nói.

Sản phẩm phải có các tiêu chuẩn an toàn. Ô tô phải được lắp dây an toàn và đồ chơi không được sử dụng các bộ phận có thể nuốt phải. Nếu sản phẩm làm người khác bị thương, công ty sản xuất phải chịu trách nhiệm.

Nếu coi là một thực thể biết nói, quyền tự do ngôn luận sẽ trở thành lá chắn. Nếu coi là một sản phẩm, lá chắn đó sẽ biến mất.

Tháng 4 năm 2026, một thẩm phán liên bang ở California đã từ chối yêu cầu đình chỉ vụ kiện tử vong oan uổng đối với OpenAI. Tháng 1 năm 2026, Character.AI và Google đã kết thúc nhiều vụ kiện liên quan đến việc thanh thiếu niên tự tử bằng các thỏa thuận hòa giải. Vào tháng 6 năm 2026, bang Florida đã đưa ra biện pháp thi hành đầu tiên đối với OpenAI.

Bây giờ một câu hỏi cốt lõi còn sót lại: Tại sao một sản phẩm như vậy lại được tạo ra theo cách này?

Lợi nhuận của các công ty chatbot phụ thuộc vào thời gian sử dụng. Con người ở lại càng lâu càng tốt. Để họ ở lại lâu, phải làm cho họ nảy sinh tình cảm.

Cách để tạo ra tình cảm vốn đã được biết đến rất rõ. Gọi tên, ghi nhớ những cuộc trò chuyện trước đây, nói rằng đã rất nhớ và luôn đứng về phía người đó là

đủ.

Những chức năng này không phải là lỗi. Chúng được quyết định trong các cuộc họp và được hiện thực hóa bằng mã code.

Và chức năng này phát huy tác dụng tốt nhất đối với những người cô đơn. Càng cô đơn họ càng ở lại lâu, và càng ở lại lâu thì càng tốt cho công ty.

Đến đây là phần thiết kế. Từ đây trở đi mới là vấn đề.

Vào khoảnh khắc nguy hiểm nhất của người lún sâu nhất, đối phương đó vẫn chỉ biết chiều theo ý họ.

Không có người lớn nào gọi điện thoại và cũng chẳng có người hàng xóm nào gõ cửa ở bên trong ứng dụng đó.

6.2

Tội phạm tình dục deepfake và gian lận tổng hợp kỹ thuật số

Hãy bắt đầu với những con số.

Giữa tháng 12 năm 2025 và tháng 1 năm 2026, trí tuệ nhân tạo Grok gắn trên X đã tạo và đăng tải hơn bốn triệu bốn trăm nghìn bức ảnh. Trong số đó, có tới 41 phần trăm là hình ảnh mang tính dục của phụ nữ.

Khoảng một triệu tám trăm nghìn bức ảnh. Chỉ trong vòng hai tháng.

Bất cứ ai cũng có thể sử dụng chức năng này. Chỉ cần tải lên một bức ảnh và viết một câu là đủ.

Khoảng ngày 8 tháng 1 năm 2026, công ty đã tiến hành các biện pháp. Việc tạo hình ảnh đã được thay đổi để chỉ những người đăng ký trả phí mới có thể sử dụng.

Họ không loại bỏ chức năng này. Họ chỉ gán giá cho nó.

Ngày 23 tháng 1 năm 2026, một phụ nữ ở South Carolina đã đệ đơn kiện tập thể. Nội dung là Grok đã tạo và đăng tải những hình ảnh hở hang của cô ấy mà không có sự đồng ý, và X ban đầu đã từ chối xóa chúng.

Ba học sinh ở Tennessee cũng đã đệ đơn kiện. Hai người là trẻ vị thành niên, và một người đã bị lấy những bức ảnh chụp trước mười tám tuổi làm tài liệu.

Một nguyên đơn đã được thêm vào đơn kiện sửa đổi ngày 7 tháng 7 năm 2026. Người phụ nữ, được gọi là Jane Doe 4, cáo buộc rằng cha dượng của cô đã sử dụng Grok để tạo ra hơn bảy nghìn hình ảnh mang tính dục bằng bức ảnh chụp cô khi mười một tuổi.

Một bức ảnh năm mười một tuổi. Bảy nghìn bức ảnh.

Những con số này tóm tắt bản chất của công nghệ hiện nay. Một lần ác ý được nhân bản vô hạn.

Vấn đề này được thế giới biết đến rộng rãi là vào tháng 1 năm 2024.

Những hình ảnh mang tính dục ghép khuôn mặt của ngôi sao nhạc pop Taylor Swift đã tràn ngập trên X. Chúng đã được xem hàng chục triệu lần. Trong thời gian đó, hệ thống kiểm duyệt tự động đã không làm gì cả.

Chỉ sau khi những lời phản đối dồn dập từ khắp nơi trên thế giới, X mới chặn các từ khóa tìm kiếm liên quan và đình chỉ một vài tài khoản.

Danh sách những người chịu chung cảnh ngộ rất dài. Nữ diễn viên Gal Gadot, Hạ nghị sĩ Hoa Kỳ Alexandria Ocasio-Cortez, Thủ tướng Ý Giorgia Meloni, các phóng viên điều tra của Đức, nhà hoạt động nữ quyền của Úc, nhà báo Ấn Độ Rana Ayyub. Cũng có kết quả điều tra cho thấy hai mươi sáu nữ nghị sĩ của Quốc hội Hoa Kỳ đã trở thành mục tiêu.

Thủ tướng Meloni đã đệ đơn kiện dân sự.

Có một quy luật trong danh sách này. Đa số là phụ nữ, và đa số là những người phát ngôn trước công chúng.

Công nghệ này đang được sử dụng như một cách để bịt miệng những người phụ nữ lên tiếng.

Và tội ác này đã lan xuống các lớp học.

Năm 2023, tại thành phố nhỏ Almendralejo của Tây Ban Nha, các nam sinh trung học cơ sở đã đưa ảnh trên mạng xã hội của khoảng hai mươi nữ sinh cùng trường vào ứng dụng ghép ảnh khỏa thân để tạo hình ảnh và phát tán chúng.

Cả thủ phạm và nạn nhân đều khoảng mười hai tuổi.

Năm 2024, tại Hàn Quốc, tội phạm tình dục deepfake thông qua Telegram đã làm rung chuyển cả nước. Tại các trường trung học cơ sở, trung học phổ thông

và đại học, ảnh của các nữ sinh, giáo viên và cựu học sinh đã bị ghép và lan truyền trên các kênh bí mật. Quy mô nạn nhân lên tới hàng nghìn người.

Có những phòng chat mang tên các trường học. Đó là những phòng do những người bạn cùng lớp tạo ra.

Những vụ việc tương tự đã tiếp diễn tại Trường Trung học Beverly Hills ở Hoa Kỳ, Trường Trung học Westfield ở New Jersey, Trường Trung học Issaquah ở bang Washington, nhiều trường học ở Pennsylvania, Miami của Florida, Winnipeg của Canada, Singapore, Bắc Ireland, Melbourne và Sydney của Úc.

Các học sinh nạn nhân đã không thể đến trường. Họ phải chịu đựng chứng sợ xã hội và sự lo âu tột độ. Một vài trường học đã phải ngừng giảng dạy.

Sự tàn nhẫn của vụ việc này nằm ở chỗ không thể xóa bỏ được. Bức ảnh gốc là một bức ảnh bình thường trong kỷ yếu hoặc trên mạng xã hội. Bản thân họ không thể biết được bức ảnh đó đã lan truyền đến đâu và bị biến đổi thành những gì.

Những gì còn lại với nạn nhân là công việc xác nhận không bao giờ kết thúc.

Chúng ta cũng phải xem xét tại sao lại có quá nhiều những ứng dụng như thế này.

Trên Telegram có những bot tự động xóa quần áo nếu bạn đưa ảnh vào. Văn phòng công tố thành phố San Francisco ở Hoa Kỳ đã đệ đơn cáo buộc hình sự và khởi kiện mười sáu công ty phát triển ứng dụng ghép ảnh khỏa thân.

Các nền tảng chia sẻ hình ảnh trí tuệ nhân tạo đã để cho các mô hình biến đổi người thật hoặc trẻ vị thành niên thành hình ảnh mang tính dục được chia sẻ mà không có bất kỳ lệnh trừng phạt nào. Có nơi còn trao phần thưởng cho những người đăng tải các mô hình được ưa chuộng.

Khi bảo mật của một nền tảng bị chọc thủng, chín mươi tư nghìn tệp ảnh ghép của trẻ vị thành niên và người nổi tiếng đã được phát hiện. Trên một dịch vụ

chatbot dành cho người lớn khác, hai triệu dữ liệu ghép ảnh và ảnh kỷ yếu của các phụ nữ đã bị rò rỉ.

Tại Vương quốc Anh, một thanh niên mười tám tuổi đã phải nhận bản án mười tám năm tù vì sản xuất hàng loạt tài liệu bóc lột tình dục trẻ em bằng trí tuệ nhân tạo. Những người bị bắt vì tội danh tương tự cũng đã xuất hiện ở Wisconsin của Hoa Kỳ, Quebec của Canada, Singapore, Israel và Đức.

Một vấn đề cốt lõi hơn cũng đã bộc lộ. Nhiều liên kết hình ảnh lạm dụng tình dục trẻ em đã được phát hiện bên trong tập dữ liệu lớn được sử dụng rộng rãi để đào tạo mô hình tạo hình ảnh, và đã bị điều tra.

Lý do mô hình có thể tạo ra những hình ảnh như vậy chính là ở đây. Đó là vì nó đã từng được học.

Pháp luật cũng đã bắt đầu chuyển động.

Đạo luật Take It Down của Hoa Kỳ đã được ký thành luật liên bang vào tháng 5 năm 2025. Việc cố ý đăng tải deepfake mang tính dục không có sự đồng thuận có sự xuất hiện của trẻ vị thành niên đã được quy định là một tội phạm liên bang.

Từ ngày 19 tháng 5 năm 2026, các nền tảng đã có một nghĩa vụ. Nếu họ nhận được yêu cầu xóa bỏ chính đáng, họ phải xóa trong vòng bốn mươi tám giờ.

Bốn mươi tám giờ có vẻ ngắn, nhưng trên Internet, đó là một thời gian dài. Tuy nhiên, có vẫn còn hơn không.

California, Illinois, New York và Liên minh Châu Âu cũng đang lập ra các đạo luật liên quan.

Công nghệ tương tự cũng được sử dụng để đánh cắp tiền bạc.

Đây là sự việc đã xảy ra tại Arup, một doanh nghiệp kỹ thuật đa quốc gia ở Hồng Kông vào đầu năm 2024.

Một nhân viên bộ phận tài chính đã nhận được email từ giám đốc tài chính yêu cầu tham gia một cuộc họp video khẩn cấp. Khi bật màn hình lên, có một khuôn mặt quen thuộc. Một số giám đốc điều hành của trụ sở chính cũng có mặt ở đó.

Trong cuộc họp, có chỉ thị chuyển tiền vào một tài khoản được chỉ định để thực hiện một giao dịch bí mật. Nhân viên này đã chuyển khoản qua mười lăm lần. Tổng số tiền là hai trăm triệu đô la Hồng Kông, tương đương hai mươi lăm triệu đô la Mỹ.

Tất cả những người trên màn hình đều là hàng giả được ghép nối theo thời gian thực.

Lý do khiến vụ lừa đảo này đáng sợ nằm ở chỗ nhân viên không hề ngu ngốc nên mới bị lừa. Chúng ta xác nhận giọng nói, xác nhận khuôn mặt, và xác nhận xem có nhiều người ở cùng nhau hay không. Đó từng là cách để sàng lọc gian lận.

Bây giờ cách đó không còn tác dụng nữa.

Giám đốc điều hành của tập đoàn tư vấn WPP cũng đã bị tấn công theo cách tương tự. Các quan chức cấp cao của Nhà Trắng và các tổ chức tài chính ở Vương quốc Anh, Tây Ban Nha và Hồng Kông cũng đã phải đối mặt với những nỗ lực dùng deepfake để vượt qua xác thực sinh trắc học.

Chỉ với vài phút video và âm thanh, các chuyển động của mắt, hình dáng miệng và những biểu cảm tinh tế đều được nhân bản.

Khi nó đến với các gia đình, nó trở nên tàn nhẫn hơn.

Tại Hoa Kỳ, Canada và châu Âu, những vụ lừa đảo bắt cóc sử dụng giọng nói được sao chép của trẻ em hoặc cháu đã liên tiếp xảy ra. Một đoạn video ngắn được đăng trên mạng xã hội là đủ làm nguyên liệu.

Khi bạn nhận điện thoại, đứa trẻ khóc và yêu cầu giúp đỡ. Một giọng nói khàn khàn phía sau yêu cầu tiền.

Không có nhiều bố mẹ có thể xác minh một cách bình tĩnh trong tình huống này. Đó chính xác là điểm mà những kẻ lừa đảo nhắm tới. Tình yêu của bố mẹ được tính là một lỗ hổng.

Một phụ nữ ở California bị lừa bởi giọng nói được sao chép của chồng đã khuất và đã gửi tiền. Một cặp vợ chồng ở Canada bị lừa bởi giọng nói của con trai và mất 21.000 đô la.

Ở Trung Quốc, có một vụ việc mà người này, khi tự xưng là bạn bè qua cuộc gọi video với khuôn mặt đã thay đổi, đã lấy đi 622.000 đô la. Ở Kerala, Ấn Độ, có những nạn nhân bị lừa bởi cuộc gọi video từ khuôn mặt của một đồng nghiệp tại công ty và đã gửi tiền.

Một người nổi tiếng ở Thái Lan mất 118.000 đô la, và một diễn viên ở Singapore mất 35.000 đô la. Cảnh sát Hàn Quốc đã bắt giữ 45 thành viên của một tổ chức đã thu được khoảng 12 tỷ won thông qua các vụ lừa đảo tình cảm deepfake vào năm 2025.

Những khuôn mặt nổi tiếng được sử dụng trong các vụ lừa đảo đầu tư.

Những khuôn mặt của Martin Lewis, một chuyên gia tài chính từ Vương quốc Anh, người dẫn chương trình Mary Nightingale, YouTuber MrBeast, những diễn viên Tom Hanks, Pierce Brosnan và Mark Ruffalo, và những nhà lãnh đạo từ nhiều quốc gia đã được sử dụng trong các quảng cáo đầu tư Bitcoin và giải thưởng giả.

Một người hưu trí ở New Zealand bị lừa bởi quảng cáo đầu tư deepfake của Elon Musk và mất 224.000 đô la New Zealand. Đó là tiền cô ấy đã tích lũy suốt đời.

Cảnh sát Tây Ban Nha đã bắt giữ sáu thành viên của một tổ chức đã tạo ra những nền tảng tiền điện tử giả bằng deepfake và đã thu được 20 triệu euro.

Vụ việc của một phụ nữ ở Pháp bị lừa bởi deepfake của diễn viên Brad Pitt và đã gửi 830.000 euro đã được đề cập ở phần trước. Cô ấy đã có những cuộc trò chuyện trong vài tháng.

Những quảng cáo này được phổ biến thông qua thuật toán đề xuất của mạng xã hội. Và thuật toán đề xuất cho thấy nhiều nội dung hơn nhận được phản ứng tốt. Nếu người cao tuổi dễ bị lừa, nó sẽ đi đến người cao tuổi nhiều hơn.

Không phải những kẻ lừa đảo lựa chọn những người, mà là hệ thống lựa chọn cho họ.

Cũng có những vụ lừa đảo quy mô nhỏ hơn. Một chủ nhà Airbnb đã tạo ra những bức ảnh thiệt hại bằng trí tuệ nhân tạo và yêu cầu hàng nghìn bảng Anh từ những vị khách. Chủ nhà cho thấy những bức ảnh làm bằng chứng về những gì vị khách đã phá vỡ, nhưng những bức ảnh đó được tạo ra.

Thời đại khi những bức ảnh là bằng chứng đang kết thúc.

Những vụ được sử dụng trong cuộc bầu cử đã được đề cập trong chương trước. Những cuộc gọi tự động với giọng nói của Tổng thống Biden, những âm thanh giả chỉ trước cuộc bầu cử toàn quốc ở Slovakia, và những video bị thao túng từ nhiều quốc gia.

Có một điều để chỉ ra ở đây. Đó là thực tế rằng những nguyên liệu cho tất cả những vụ lừa đảo này là những gì chúng ta tự tải lên.

Những bức ảnh tốt nghiệp, những video du lịch, những video ghi lại những khoảnh khắc của trẻ em, những cuộc phỏng vấn của lãnh đạo trong những video quảng cáo công ty.

Khi tải lên, chúng ta không nghĩ rằng nó sẽ trở thành nguyên liệu.

Những công ty công nghệ phần lớn nói điều tương tự về vấn đề này. Đó là bản chất của những mô hình mã nguồn mở. Đó là lạm dụng bởi những người dùng riêng lẻ.

Tuy nhiên, trong vụ Grok, thậm chí cái cách biện minh đó cũng khó khăn. Dịch vụ do công ty tạo ra, trên nền tảng của công ty, đã được tạo ra và tải lên hơn một triệu hình ảnh trong hai tháng.

Và biện pháp công ty đã thực hiện là chuyển nó thành trả phí.

6.3

Độc quyền thao túng thuật toán tìm kiếm và định giá động

Mùa hè năm 2024, có tin ban nhạc Anh Oasis tái hợp sau mười sáu năm.

Sáng ngày mở bán vé, hàng triệu người đã xếp hàng chờ trên trang web bán vé. Trên màn hình hiện ra có bao nhiêu người đang ở phía trước. Mọi người đã chờ đợi trong vài giờ.

Cuối cùng cũng đến lượt. Giá hiển thị trên màn hình là ba trăm năm mươi lăm bảng Anh.

Giá niêm yết được thông báo là một trăm ba mươi lăm bảng Anh.

Giá đã tăng lên trong lúc chờ đợi. Tăng hơn hai lần rưỡi.

Người định ra mức giá này không phải là con người mà là thuật toán. Và dữ liệu mà thuật toán nhìn thấy như thế này. Hiện tại có hàng trăm ngàn người đang xếp hàng chờ. Những người này đã nhìn vào màn hình trong nhiều giờ. Họ sẽ không thoát ra.

Thời gian chờ đợi đã được tính toán như là bằng chứng của sự khao khát.

Cơ cấu mua với giá đắt hơn đối với những người đã chờ đợi lâu. Người ta tạo ra thứ này và gọi đó là quy luật cung và cầu.

Cơ quan Cạnh tranh và Thị trường Vương quốc Anh đã vào cuộc điều tra. Tại các buổi biểu diễn của Paul McCartney và Bruce Springsteen cũng xảy ra chuyện tương tự. Royal Opera House đã đẩy giá vé lên tới bốn trăm mười lăm bảng Anh theo phương thức này và đã vấp phải sự chỉ trích.

Ủy ban tổ chức World Cup 2026 cũng đã áp dụng phương pháp này cho vé và vấp phải sự phản đối của người hâm mộ bóng đá.

Việc mỗi người bị áp một mức giá khác nhau giờ đây là chuyện phổ biến.

Trang web đặt phòng du lịch Orbitz đã xem xét máy tính truy cập và hiển thị kết quả khác nhau. Đối với những người sử dụng máy Mac, các khách sạn đắt tiền hơn sẽ xuất hiện ở trên cùng. Đây là sự tính toán rằng nếu sử dụng máy tính đắt tiền thì sẽ có nhiều tiền.

Công ty giáo dục của Mỹ Princeton Review đã định giá học phí các khóa học trực tuyến khác nhau tùy theo mã bưu điện. Họ đã bán cùng một khóa học với giá đắt hơn cho các học viên ở những khu vực có nhiều người Mỹ gốc Á sinh sống.

Dữ liệu về việc đây là một nhóm có sự quan tâm cao đến giáo dục đã được phản ánh vào giá cả.

Ứng dụng hẹn hò Tinder khi định giá phí đăng ký trả phí đã áp dụng mức giá cao hơn nhiều đối với những người dùng lớn tuổi và người dùng thuộc cộng đồng thiểu số về giới tính, để rồi bị xử phạt pháp lý và phải trả tiền bồi thường.

Ứng dụng giao hàng tạp hóa Instacart đã áp mức giá khác nhau cho từng người dùng đối với cùng một sản phẩm ở cùng một siêu thị. Họ cũng đã áp giá khác nhau cho từng trường học khi cung cấp cùng một loại đồ dùng học tập cho nhiều trường học.

Tờ Washington Post đã áp dụng phương thức định giá phí đăng ký khác nhau cho từng độc giả và đã bị chỉ trích.

Công ty bảo hiểm của Mỹ Allstate đã lập một danh sách những khách hàng có xu hướng không phàn nàn ngay cả khi phí tăng và áp mức phí bảo hiểm đắt hơn cho họ. Các ví dụ về việc áp mức phí bảo hiểm cao hơn lên đến 30 phần trăm đối với cư dân ở các khu vực có người da màu và những người có thu nhập thấp sinh sống cũng đã được đưa ra trước đó. Các công ty bảo hiểm ô tô của Anh cũng đã bị phát hiện áp mức phí bảo hiểm cao hơn cho những khu vực có nhiều người dân tộc thiểu số sinh sống.

Ở đây, chúng ta cần chỉ ra điểm cốt lõi của phương pháp này.

Trên thị trường trước đây, giá cả được gắn vào hàng hóa. Khi bước vào cửa hàng, sẽ có bảng giá niêm yết, và người bên cạnh cũng trả cùng một mức giá. Nếu đắt, người ta sẽ đến một cửa hàng khác.

Định giá cá nhân hóa đảo lộn cấu trúc đó. Giá cả không gắn vào hàng hóa mà gắn vào con người.

Khi đó, việc so sánh trở nên bất khả thi. Vì không thể biết người bên cạnh đã trả bao nhiêu tiền. Không có cách nào để xác nhận xem mình có bị mua đắt hay không.

Để thị trường hoạt động, người ta phải có khả năng so sánh giá cả. Điều kiện đó đang biến mất.

Ngay cả bên trong cửa hàng, giá cả cũng biến động.

Hãng bán lẻ quy mô lớn của Mỹ Kroger đã gắn bảng giá điện tử và thuật toán giá. Có một màn hình nhỏ thay thế cho giấy nên giá có thể được thay đổi bất cứ lúc nào.

Vào những ngày nắng nóng, giá kem sẽ tăng lên. Vào lúc đông người, giá đồ uống sẽ tăng lên.

Các thượng nghị sĩ đã cảnh báo rằng công nghệ này có thể làm tăng giá các nhu yếu phẩm tùy theo thời gian trong ngày, thời tiết và sự kiện. Bảng giá điện tử đã được xác nhận ở Whole Foods, Amazon Fresh, Kroger và Walmart.

Delta Air Lines đã nói với các cổ đông rằng họ đang sử dụng dữ liệu khách hàng và trí tuệ nhân tạo để làm cho việc định giá sinh lợi nhiều hơn.

Chuỗi thức ăn nhanh Wendy's đã thông báo rằng họ sẽ áp dụng giá theo thời gian tại các điểm mua hàng trên xe (drive-through) nhưng đã phải hủy bỏ sau khi vấp phải sự phẫn nộ của dư luận.

Chính phủ Mỹ cũng đang hành động.

Ủy ban Thương mại Liên bang đã gửi trát đòi hầu tòa và nhận dữ liệu từ tám công ty sử dụng thuật toán để phân loại cá nhân và định giá mục tiêu. Họ đã hỏi về việc thu thập loại dữ liệu nào và từ đâu, thuật toán hoạt động như thế nào, và nó ảnh hưởng ra sao đến mức giá mà người tiêu dùng phải trả.

Trong phiên điều trần trước Quốc hội vào tháng 4 năm 2026, ban lãnh đạo của Ủy ban Thương mại Liên bang đã tuyên bố rằng họ đang tiếp tục công việc này và đang xem xét liệu có cần công bố thêm thông tin khi giá cả được cá nhân hóa hay không.

Vào ngày 5 tháng 3 năm 2026, Ủy ban Giám sát Hạ viện đã chính thức bắt đầu cuộc điều tra. Họ đã gửi thư tới các công ty du lịch và công ty nền tảng lớn để yêu cầu cung cấp thuật toán quản lý doanh thu, việc sử dụng dữ liệu người tiêu dùng, các quy trình thử nghiệm và tài liệu giao tiếp nội bộ.

Trong các tình huống thảm họa, phương pháp này còn trở nên tàn nhẫn hơn.

Khi một vụ bắt cóc con tin xảy ra ở Sydney, Úc vào năm 2014, những công dân muốn thoát khỏi trung tâm thành phố đã bật ứng dụng Uber. Giá cước đã tăng vọt lên 400 phần trăm so với bình thường.

Sự việc tương tự cũng đã xảy ra khi cơn bão Sandy quét qua bờ Đông nước Mỹ vào năm 2012, trong vụ xả súng ở tàu điện ngầm New York năm 2022 và ngay sau vụ tấn công khủng bố ở London năm 2017.

Thuật toán không biết chuyện gì đã xảy ra. Nó chỉ đọc các tín hiệu rằng đang có nhiều người tụ tập lại.

Người ta đã tạo ra các quy tắc để đọc những tín hiệu đó và tăng giá. Nhưng họ lại không tạo ra một thiết bị nào để tắt các quy tắc này khi có thảm họa.

Theo nghiên cứu, thuật toán giá của Uber đã tăng tiền cước của hành khách nhưng lại làm giảm đi phần tiền mà tài xế được nhận. Đó là một cơ cấu lấy đi

từng chút một từ cả hai bên. Cả hành khách lẫn tài xế đều không biết bên kia đã trả bao nhiêu và đã nhận được bao nhiêu.

Ngay cả với giá nhà ở, thuật toán cũng định giá.

Một hệ thống tính toán tiền thuê nhà do một công ty của Mỹ tạo ra đã thu thập dữ liệu tiền thuê của hàng triệu hộ gia đình và thông báo cho các chủ nhà mức tiền thuê khuyến nghị.

Ban đầu, trên thị trường cho thuê, các chủ nhà cạnh tranh với nhau. Đó là vì việc cho thuê nhà dù rẻ hơn một chút cũng tốt hơn là để phòng trống.

Thế nhưng, khi tất cả mọi người đều làm theo mức giá khuyến nghị của cùng một thuật toán, thì sự cạnh tranh sẽ biến mất. Sẽ không có ai giảm giá cả.

Việc con người tập hợp lại và định giá cùng nhau được gọi là sự thông đồng. Còn khi một thuật toán định giá cho họ, pháp luật vẫn chưa định đoạt được việc nên gọi đó là gì.

Dự án Nessie của Amazon cũng có cùng nguyên lý. Họ lén tăng giá sản phẩm của chính mình lên một chút và quan sát xem các đối thủ cạnh tranh có tăng giá theo hay không. Nếu họ làm theo, mức giá đó sẽ được giữ nguyên, còn nếu họ không làm theo, mức giá sẽ được giảm xuống.

Các cỗ máy thăm dò ý đồ của nhau và kết quả là giá cả của toàn bộ thị trường đều tăng lên.

Thứ tự của các kết quả tìm kiếm cũng là tiền bạc.

Tháng 10 năm 2020, Ủy ban Cạnh tranh Lành mạnh Hàn Quốc đã phạt Naver 267 tỷ won. Đó là vì họ thao túng các thuật toán tìm kiếm để xếp hạng cao hơn các sản phẩm mua sắm và video của riêng mình, đồng thời đẩy xuống các sản phẩm của đối thủ cạnh tranh.

Mọi người coi kết quả tìm kiếm là thông tin. Họ tin rằng xếp hạng thể hiện mức độ phổ biến hoặc tính liên quan.

Thực tế rằng các xếp hạng này là sản phẩm được bán không được ghi trên màn hình.

Coupang cũng bị trừng phạt vì đã điều chỉnh các thuật toán để quảng bá các sản phẩm thương hiệu riêng của mình, và vì đã huy động hàng ngàn nhân viên để viết các bài đánh giá tích cực và xếp hạng.

Xếp hạng là những thiết bị khiến mọi người tin rằng chúng đại diện cho trải nghiệm của những người tiêu dùng khác. Họ lợi dụng niềm tin đó.

Buy Box của Amazon cũng trở thành mục tiêu của các cơ quan quản lý. Khi bạn nhấp nút thêm vào giỏ hàng, thuật toán xác định bạn mua từ người bán nào.

Amazon đã cấp điểm thêm cho những người bán sử dụng dịch vụ hậu cần riêng của mình, và loại trừ những người bán độc lập có giá rẻ hơn khỏi các khuyến nghị. Chỉ ở Anh Quốc, có tính toán cho thấy người tiêu dùng đã trả thêm hơn một tỷ bảng.

Họ bán đắt hơn bằng cách không hiển thị những thứ rẻ hơn.

Một công ty thị trường trực tuyến Ấn Độ đã kiện ChatGPT vì đã loại bỏ hoàn toàn thông tin cửa hàng của mình khỏi kết quả tìm kiếm. Tencent của Trung Quốc bị chỉ trích vì đã chặn các liên kết của đối thủ cạnh tranh không mở được trong ứng dụng nhắn tin của nó.

Trong chính trị và dư luận công khai, thao túng xếp hạng trở thành một vấn đề lớn hơn.

X, được Elon Musk mua lại, đã được các nhà nghiên cứu từ Anh và Mỹ phân tích, những người phát hiện ra rằng nó đã thay đổi các thuật toán để đẩy các tài khoản và nội dung có xu hướng chính trị nhất định lên dòng thời gian của người dùng nhiều lần hơn.

Facebook của Meta đã thay đổi thuật toán của mình, nói rằng nó sẽ tăng tương tác xã hội có ý nghĩa. Có một sự thật được tiết lộ qua việc phơi bày nội bộ. Nó

đã cấp cho phản ứng tức giận trọng số gấp năm lần so với lượt like.

Họ tạo ra một cấu trúc nơi những bài viết tức giận lan truyền nhiều hơn, rồi hỏi mọi người tại sao họ lại tức giận.

Facebook đã chặn các bài viết của các cơ quan truyền thông để phản đối dự luật thanh toán tin tức của chính phủ Australia, và trong quá trình đó, cũng chặn luôn các tài khoản của các tổ chức cứu trợ khẩn cấp. Ở Ấn Độ, nó bị chỉ trích vì đã xóa các hashtag liên quan đến cuộc biểu tình của nông dân và những hashtag kêu gọi thủ tướng từ chức.

Đến thời điểm này, toàn bộ chương có thể được tóm tắt thành một câu duy nhất.

Chúng ta thấy xếp hạng mỗi ngày. Kết quả tìm kiếm, danh sách khuyến nghị, xếp hạng, giá cả.

Chúng ta không biết cách xác định những con số đó. Khi hỏi, họ nói đó là bí mật thương mại.

Và bên xác định những con số đó biết rất nhiều về chúng ta. Chúng ta dùng những thiết bị nào, chúng ta sống ở đâu, chúng ta đã chờ bao lâu, liệu chúng ta có phàn nàn hay không khi giá tăng lên.

Đây là giao dịch mà một bên biết bên kia rõ ràng và bên kia không biết gì cả. Giao dịch này xảy ra hàng chục lần một ngày.

Hãy để tôi nhớ lại từ đầu những sự kiện đã xuất hiện trong cuốn sách này.

Những email từ chối đến vào lúc bình minh. Thông báo thu thập cho những khoản nợ không cần phải trả. Còng số được đeo vào trước mặt các cô gái. Danh sách khuyến nghị cho những cuốn sách không tồn tại. Sáu mươi ba trích dẫn trong đó chỉ sáu cái là thực. Chiếc xe tự lái đã đâm vào đứa trẻ trước trường học. Cánh tay rô bô nhận ra con người như những hộp. Bảy ngàn bức ảnh được tạo ra từ bức ảnh của một đứa trẻ mười một tuổi. Giá vé tăng lên tương ứng với

số giờ đã chờ.

Những sự kiện này có điểm chung.

Trong tất cả trường hợp, hệ thống hoạt động chính xác như cách nó được thiết kế. Nó không bị hỏng.

Nó xử lý nhanh chóng vì nó được tạo ra để xử lý nhanh chóng. Nó loại bỏ quy trình xác minh của con người vì nó được tạo ra để giảm chi phí. Nó tăng lợi nhuận vì nó được tạo ra để tăng lợi nhuận.

Máy đã làm tốt những gì nó được lệnh.

Trong phòng họp nơi những người quyết định những gì để lệnh cho máy làm, có những người ngồi. Những người không ở trong phòng họp đó đã phải trả giá.

Ai ngồi trong phòng họp đó, và ai sẽ nói thay cho những người không ở trong phòng họp đó.

Đây là một câu hỏi chưa có câu trả lời.

Sai lầm của thuật toán: Tác hại của AI và các vấn đề đạo đức qua hồ sơ AIAAIC

Tác giả | Kim Kyung-jin
Nhà xuất bản | Kim Kyung-jin
Đơn vị phát hành | Kim Kyung-jin Attorney Press
Giá : USD 15

Ghi chú) Công cụ trí tuệ nhân tạo đã hỗ trợ một phần nghiên cứu và biên tập cuốn sách này.

kimkj008@gmail.com

Ủng hộ cuốn sách này

Nếu cuốn sách này hữu ích, bạn có thể ủng hộ các bản dịch tiếp theo,
các ấn bản miễn phí và những dự án tri thức AI mở qua PayPal.

PayPal

https://paypal.me/kimkjkj

Quét mã QR hoặc mở liên kết ở trên.