NHẬN DIỆN THÔNG TIN SAI LỆCH TỪ NỘI DUNG AI TẠO: VẤN ĐỀ VÀ GIẢI PHÁP

Phan Bá Hải Triều
Trường Đại học Hòa Bình
Tác giả liên hệ: pbhtrieu@daihochoabinh.edu.vn

Ngày nhận: 12/9/2025
Ngày nhận bản sửa: 20/10/2025
Ngày duyệt đăng: 24/10/2025
DOI: 10.71192/655737yktyhf

Tóm tắt

Sự tiến bộ nhanh chóng của các công cụ trí tuệ nhân tạo tạo sinh đã mang lại những trải nghiệm đáng kinh ngạc trong "thế giới" truyền thông, song cũng đặt ra những thách thức nghiêm trọng về thông tin sai lệch và tin giả, đặc biệt trong bối cảnh tình hình địa chính trị trên thế giới diễn ra ngày một phức tạp. Các nội dung đa phương tiện siêu thực do AI tạo ra (mà các mô hình như Google Veo3, KlingAI là ví dụ điển hình) đã làm thay đổi cục diện và bản chất của vấn đề, dễ dàng bóp méo sự thật và tạo ra một "thực tế tổng hợp" khó phân biệt với sự kiện, nhân vật thực tế. Sự hào hứng và giải trí ban đầu do các mô hình AI đem đến cho người dùng nhanh chóng trở thành mối nguy đáng báo động, dễ bị lạm dụng hơn bao giờ hết. Bài viết này chỉ ra những đặc điểm và nghiên cứu mối đe dọa của thông tin sai lệch tạo bởi AI, đồng thời, phân tích các phương pháp phát hiện dựa trên logic thực tế, dựa trên AI, thủy vân số, sự phối hợp của cộng đồng. Ngoài ra, bài viết cũng đề xuất một số phương cách giảm thiểu tác động của tin giả, tin sai lệch tạo bởi AI để bảo vệ tính toàn vẹn của thông tin và niềm tin xã hội trong kỷ nguyên số.

Từ khóa: *Tin giả do AI, Deepfake, Veo3, phát hiện AI, trách nhiệm nền tảng.*

Detecting AI-Generated Misinformation: Challenges and Solutions

Phan Ba Hai Trieu
Hoa Binh University
Corresponding Author: pbhtrieu@daihochoabinh.edu.vn

Abstract

The rapid advancement of generative artificial intelligence tools has brought astonishing experiences to the media "world," but it has also posed serious challenges concerning misinformation and fake news—especially amid increasingly complex global geopolitical dynamics. Advanced AI models such as Google Veo3 and KlingAI have ushered in a new era where hyper-realistic multimedia content blurs the line between truth and fiction, giving rise to a "synthetic reality" that is increasingly difficult to discern from actual events and individuals. What began as a source of excitement and amusement has quickly transformed into a concerning risk, heightened by the potential for misuse. This paper identifies the characteristics and examines the threats of AI-generated misinformation, while analyzing detection methods based on factual logic, AI-driven approaches, digital watermarking, and community collaboration. In addition, it proposes several measures to mitigate the impact of AI-generated fake and misleading information, in order to safeguard the integrity of information and maintain social trust in the digital era.

Keywords: *AI-generated Fake News, Deepfake, Veo3, AI detection, Platform Responsibility.*

1. Đặt vấn đề

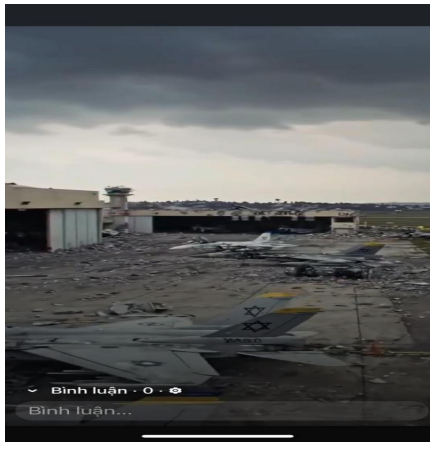
Là "điện mới của nhân loại" (GS. Andrew Ng, Đại học Stanford, 2017), trí tuệ nhân tạo, hay còn gọi Artificial Intelligence (AI) đã nhanh chóng trở thành công cụ vạn năng, vượt qua nhiều giới hạn từng được biết tới và đang trên đà trở thành xu hướng công nghệ chủ đạo của Cách mạng công nghiệp 4.0. AI tác động đến mọi lĩnh vực, và báo chí - truyền thông ở Việt Nam không phải là một ngoại lệ. Dẫu có nhiều ưu điểm, là một trợ thủ đắc lực, cung cấp khả năng sáng tạo, tổng hợp không giới hạn, song ở chiều kích khác, AI lại đang làm thay đổi cục diện và bản chất của vấn đề tin giả - tin sai lệch vốn tồn tại cố hữu trên môi trường truyền thông nhiều năm nay. Những nội dung đa phương tiện do AI tạo ngày càng thực tế, thậm chí đạt đến mức độ siêu thực, khó nhận diện, dễ dàng đánh

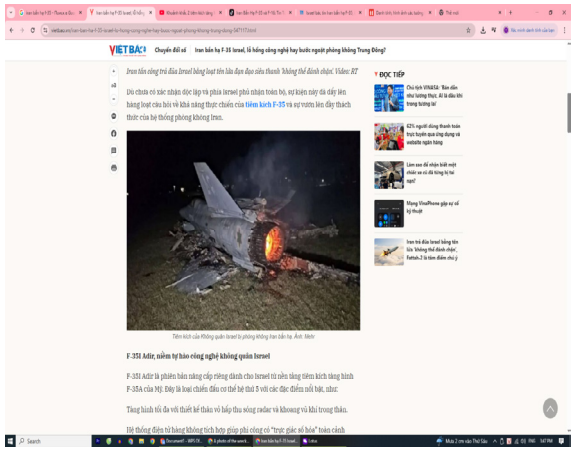
lừa công chúng và các chuyên gia. Trong bối cảnh bùng nổ thông tin, cuộc chiến với thông tin sai lệch (misinformation) và tin giả (disinformation) tạo bởi AI giờ đây đã trở thành thách thức toàn cầu [1].

Ước tính, có khoảng 500.000 video deepfake tạo bởi AI được chia sẻ trên mạng xã hội vào năm 2023. Trong khi ở Việt Nam, số tiền thiệt hại vì lừa đảo trực tuyến ước lượng lên tới 18.900 tỷ đồng, nhiều trường hợp trong đó, tội phạm sử dụng AI để gia tăng tỷ lệ thành công khi tiếp xúc với đối tượng tiềm năng. Theo báo cáo của Exploding Topics, đến năm 2025, ước tính có 8,4 tỷ trợ lý ảo được sử dụng trên toàn thế giới. Số lượng nội dung đa phương tiện siêu thực tạo bởi AI ngày một tăng cao. Chỉ trong một năm từ 2022 tới 2023, đã có hơn 15 tỷ bức ảnh được tạo ra bởi thuật toán AI nhờ công nghệ text-to-image. Nhiều bức ảnh trong số đó đã đánh lừa được các tòa soạn và trang tin điện tử trong nước.

Ví dụ, vào rạng sáng ngày 14/6/2025 (giờ Hà Nội), sau cuộc tấn công của Iran vào thủ đô Tel Aviv của Israel, trên mạng xã hội và nhiều phương tiện truyền thông đại chúng (trong đó, có một số tờ báo điện tử, trang thông tin điện tử của Việt Nam) đã đăng tải một số nội dung về chiến đấu cơ F-35, sân bay quân sự Israel bị phá hủy. Các nội dung video, hình ảnh mang chi tiết chân thực khó phân biệt, ánh sáng hài hòa, phối cảnh xa gần tiệm cận ở mức độ hoàn hảo. Song nếu nhìn thật kỹ vào các chi tiết, sẽ thấy một số điểm bất hợp lý, như biểu tượng insignia của không quân, cánh đuôi lắp ngược, hệ thống phản lực còn nóng đỏ dù máy bay đã bị hạ. Sự phi logic này đã dẫn nhiều phân tích đến kết luận: Không loại trừ khả năng, đây là các nội dung đa phương tiện siêu thực được tạo bởi các mô hình AI tạo sinh.

Hình 1 và Hình 2. Video sân bay quân sự Israel trên Youtube và hình ảnh máy bay có biểu tượng nhà nước Do Thái bị bắn hạ





Nguồn: vietbao.vn/iran-ban-ha-f-35-israel-lo-hong-cong-nghe-hay-buoc-ngoat-phong-khong-trung-dong-547117.htm

Trong bài viết này, thông tin sai lệch được định nghĩa là thông tin không chính xác hoặc gây hiểu lầm được lan truyền mà không có ý định lừa dối, trong khi tin giả là thông tin sai lệch được tạo ra và phát tán có chủ đích nhằm mục đích lừa gạt, thao túng hoặc gây hại [1].

Thuật ngữ AI tạo sinh - còn gọi là Generative AI (GenAI) - là một nhánh của trí tuệ nhân tạo, có khả năng tạo ra nội dung mới, độc đáo dựa trên dữ liệu mà nó đã được huấn luyện, nổi bật với khả năng "sáng tạo" [2].

Trước đây, tin giả truyền thống, thường được gọi là "tin tức giả" (fake news), chủ yếu là một hình thức báo chí lá cải hoặc truyền thông nhằm mục đích phân phối thông tin sai lệch có chủ đích thông qua các phương tiện in ấn hoặc mạng xã hội trực tuyến [3]. Loại tin tức này đã trở thành một vấn đề lớn với các cá nhân nổi tiếng và nhiều quốc gia, tuy nhiên, sự xuất hiện của AI tạo sinh đã đánh dấu một sự chuyển đổi cơ bản. Thay vì chỉ làm giả thông tin hoặc bóp méo sự thật, AI giờ đây có khả năng "tạo ra thực tế" hoàn toàn mới, siêu thực và khó phân biệt. Hiện nay, GenAI được tích hợp vào đủ ứng dụng, từ Google, Facebook, Tiktok cho tới Zalo,… và có thể hoạt động ngay trên cả các thiết bị di động cấu hình thấp, giá rẻ nhờ công nghệ đám mây tân tiến.

Deepfake - ứng dụng gây tranh cãi nhất của GenAI, đại diện cho xu hướng mới nhất trong thông tin sai lệch, sử dụng AI để tạo ra các video giả mạo [3]. Thuật ngữ "deepfake" là sự kết hợp của "deep learning" (học sâu) và "fake" (giả mạo), miêu tả kỹ thuật tổng hợp hình ảnh người, vật dựa trên AI. Sự chuyển dịch từ "tin tức giả" sang "thực tế tổng hợp" này đặt ra dấu hỏi lớn cho các nhà truyền thông trong quá trình nhận thức và xác minh thông tin, đòi hỏi các phương pháp phát hiện và tư duy phản biện vượt ra ngoài cách xác thực thông thường. Hiện nay, sự phát triển vũ bão của các mô hình AI tạo sinh đã khuếch đại đáng kể tốc độ và sự tinh vi của việc lan truyền tin giả, khiến việc xác định và chống lại thông tin sai lệch ngày càng khó khăn [4]. Diễn đàn Kinh tế Thế giới đã xác định thông tin sai lệch và tin giả là mối đe dọa lớn nhất đối với nhân loại trong hai năm tới và là rủi ro toàn cầu đáng kể thứ năm về lâu dài [1].

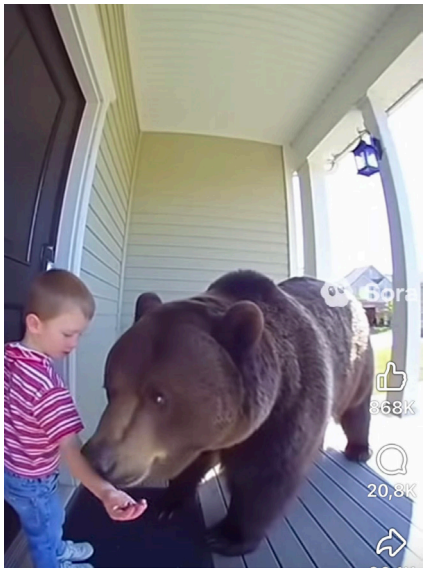
2. Thách thức và đặc điểm do thông tin sai lệch do AI tạo

2.1. Thách thức với thông tin sai lệch do AI tạo

2.1.1. Độ chân thực vượt quá khả năng xác minh của công chúng

Với sự ra mắt của một số AI tạo ảnh, video như Veo3, Nano Banana (Google) hay Sora 2 (Open AI), nhiều người dùng và dư luận trên mạng xã hội đã bày tỏ lo ngại về mức độ chân thực quá mức của các mô hình này, gây nhầm lẫn nghiêm trọng về nguồn gốc thực tế.

Hình 3. Video được tạo bởi mô hình Sora 2 thu hút 868.000 lượt thích trên mạng xã hội Tiktok và gây nhầm lẫn cho nhiều chủ tài khoản



Nguồn: Tiktok

Với tần suất và thói quen tiêu thụ nội dung ngày càng ngắn và nhanh của người dùng mạng, của sinh viên, học sinh và khả năng phân biệt bằng mắt thường kém đối với người cao tuổi, việc phân định thật giả càng trở nên khó khăn. Thông tin siêu thực do AI tạo cũng được cá nhân hoá theo thuật toán, khiến những người nhầm lẫn với thông tin sai lệch ngày càng lún sâu vào các phân cực và vòng lặp nội dung tạo bởi trí tuệ nhân tạo.

2.1.2. Tốc độ lan truyền nhanh và làm xói mòn niềm tin của xã hội

Một nghịch lý đáng chú ý là mặc dù phần lớn công chúng (91,9%) đã nhận thức được AI và lo ngại về việc nó tạo ra tin giả (86%) [4], thông tin sai lệch do AI lại thường tập trung vào nội dung giải trí và có khả năng lan truyền mạnh mẽ hơn đáng kể so với các hình thức thông tin sai lệch thông thường [5]. Điều này tạo ra một thách thức về nhận thức: người dùng có thể nhận thức được mối đe dọa chung, nhưng lại dễ bị cuốn hút và chia sẻ nội dung do AI tạo ra do tính chất giải trí và cảm xúc tích cực của nó, làm mờ đi ranh giới giữa giải trí và thông tin, từ đó, bình thường hóa việc tiếp xúc với nội dung giả mạo. Việc ngày càng nhiều người sử dụng một số mô hình AI tạo video như RunwayML, KlingAI,… đã khiến công chúng và cả những người làm truyền thông ngày càng bớt cảnh giác

với nội dung tạo bởi AI. Từ đó, một sự hoài nghi toàn diện diễn ra ngay cả với các nội dung văn bản, phỏng vấn của cơ quan chức năng, báo chí chính thống.

Trong cuốn sách *"Calling Bullshit: The Art of Skepticism in a Data-Driven World"*, hai giáo sư Carl T. Bergstrom và Jevin D. West (Đại học Washington (Mỹ)) đã nhắc đến khái niệm "Cổ tức của kẻ nói dối" (Liar's Dividend). Khái niệm chỉ ra mối đe dọa tiềm ẩn nghiêm trọng nhất từ AI tạo sinh là việc không chỉ tạo ra tin giả mà còn làm suy yếu khả năng của xã hội trong việc phân biệt sự thật khỏi giả dối. Tác giả nhận xét: *"... ngay cả khi một video là thật và bạn không có lý do gì để nghi ngờ, (nhưng bởi AI), bạn vẫn phải đặt câu hỏi liệu nó có thật hay không?"* [27]. Thách thức này đòi hỏi các giải pháp không chỉ về công nghệ mà còn về cả giáo dục đội ngũ công chúng và đội ngũ làm truyền thông thiết lập tư duy phản biện đối với tất cả tin tức hiện nay [20]. Trong thời điểm hiện tại, bằng nhiều phương thức và tận dụng đặc điểm, các nội dung AI vẫn có thể nhận diện với người không có chuyên môn về kỹ thuật.

2.2. Đặc điểm của thông tin sai lệch do AI tạo ra

Với AI tạo sinh, một người có thể tạo ra nhiều loại hình tin tức giả mạo. Tin giả tạo bởi AI được phân loại là:

- Deepfake (đối với hình ảnh/video/ âm thanh)

- Văn bản giả mạo

Bên cạnh đó, AI cũng có thể "bịa đặt" nội dung trong "vô thức" (còn gọi là hiện tượng "ảo giác AI"), biến một thông tin trở nên sai lệch về một số chi tiết, gây nhầm lẫn cho người dùng khi chia sẻ. Về bản chất, các mô hình ngôn ngữ lớn LLM như Gemini, ChatGPT hoạt động dựa trên việc ghi nhớ, phân loại dữ liệu khổng lồ rồi dự đoán từ ngữ tiếp theo bằng các thuật toán, do vậy, nếu như không có "mẫu", thì AI như một người mù mò mẫm trong bóng tối mà không thật sự có tư duy như con người.

Nhìn chung, các loại tin giả, tin sai lệch tạo bởi AI sẽ có các đặc điểm chung:

- Tính thuyết phục cao (do mức độ chân thực đáng kinh ngạc).

- Tốc độ tạo lập cao, phạm vi lan truyền rộng (tự động hóa và mở rộng).

- Nội dung gây sốc, kích động (để khuyến khích chia sẻ).

- Có khả năng cá nhân hóa (theo sở thích, vị trí địa lý của người dùng).

- Khó truy nguồn gốc (do bối cảnh, nhân vật, sự kiện đều không có thực).

Trong các loại tin giả, deepfake là một trong những ứng dụng AI tạo sinh gây lo ngại nhất, bởi chi tiết được mô phỏng tối đa từ thực tế, đặc biệt là khuôn mặt và giọng nói con người. Đây là một kỹ thuật tổng hợp hình ảnh dựa trên AI, sử dụng mạng đối kháng tạo sinh (GAN) để kết hợp và chồng ghép các hình ảnh và video hiện có lên các hình ảnh hoặc video nguồn [3]. Các loại deepfake chính bao gồm hoán đổi khuôn mặt (face-swap), đồng bộ hóa môi (lip-sync), và tái tạo khuôn mặt theo thời gian thực (real-time facial reenactment) [6]. Deepfake có thể giả mạo hình ảnh phức tạp trong video, nhưng có thể chỉ để giả mạo tiếng nói của một người. Ví dụ, như cuộc gọi tự động deepfake của cựu Tổng thống Mỹ Joe Biden nhắm vào cử tri Mỹ vào hồi năm 2024 là một điển hình trong việc sử dụng AI để tác động lên kết quả chính trị [7].

Hình 4. Bản tin giả tạo bởi mô hình AI Veo3

Nhìn chung, với tính chất chân thực, so với các tin giả truyền thống, tin giả như video tạo bởi AI tạo sinh làm ảnh hưởng sâu sắc đến niềm tin của công chúng, đặc biệt đối với nhóm công chúng trung niên hoặc lớn tuổi, những người ít thông thạo hơn về công nghệ và dành không nhiều thời gian quan tâm tới lĩnh vực này. Những công chúng ít thông thạo về AI cũng có thể tạo ra một vòng lặp nguy hiểm: hào hứng với các ứng dụng AI dễ sử dụng, rồi vô tình lan truyền những thông tin sai sự thật trong vô thức.

Sự phát triển của các công cụ AI tạo sinh hiện đại như Veo3, RunwayML và Sora đã hạ thấp đáng kể cả rào cản tài chính và kỹ thuật cho việc tạo ra phương tiện tổng hợp, giúp việc tạo deepfake trở nên dễ dàng và chất lượng cao hơn [8]. Các công cụ này đã "thương mại hóa" công nghệ deepfake, làm giảm "rào cản tài chính và kỹ thuật" cho việc sử dụng chúng, khả năng tạo deepfake không còn giới hạn ở các chuyên gia kỹ thuật cao mà đã trở nên phổ biến, dễ tiếp cận, dẫn đến một "lũ lụt" nội dung deepfake trên internet [8]. Ví dụ, phí thuê bao để tạo nội dung video với Veo3 chỉ khoảng 500.000 Vnđ/ tháng. Trong khi đó, với quan niệm "chỉ để giải trí", người dùng thường vô tư chia sẻ các nội dung sai lệch tạo bởi AI, khiến các nội dung giả mạo tạo bởi AI có thêm một đặc điểm: phổ biến và lan rộng với các tài khoản mạng xã hội nhỏ, ẩn danh, làm gia tăng thêm thách thức khi truy vết khởi nguồn của tin tức.

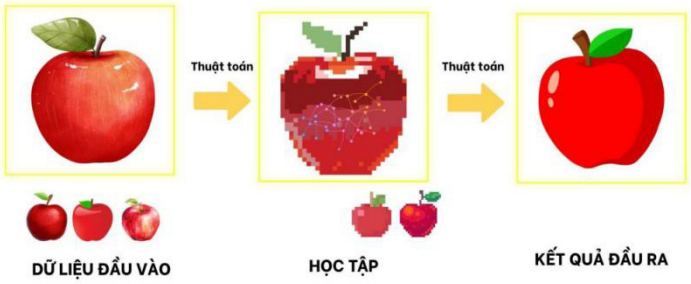
3. Phương pháp nhận diện thông tin sai lệch, tin giả tạo bởi AI

Dù so với chính mình cách đây một vài năm, AI tạo sinh đã đạt những bước tiến dài, ngày càng tinh vi, chi tiết, song tin giả tạo bởi AI không phải là không thể nhận diện trong bối cảnh hiện nay (dù một vài phương pháp nhận diện chỉ có ý nghĩa tính bằng ngày giờ do AI có khả năng tự học hỏi và tự huấn luyện ưu việt).

3.1. Nhận diện thông qua logic, quan sát

Đối với AI tạo sinh, các thuật toán đóng vai trò là tập hợp các quy tắc và hướng dẫn để AI xử lý thông tin, tạo ra các cấu trúc toán học để các mô hình AI phân tích và tìm kiếm mối liên hệ rồi đưa ra quyết định. Ví dụ, để sáng tạo ra hình ảnh một quả táo, AI tạo sinh như ChatGPT thực chất không có ý niệm về quả táo như bộ não của con người, mà có thể sử dụng phương án ghi nhớ các cách sắp xếp pixel *(hay còn gọi là "picture element" (phần tử hình ảnh), về bản chất là đơn vị cơ bản nhỏ nhất để tạo nên một hình ảnh kỹ thuật số hoặc hiển thị trên màn hình)* trong hình ảnh của một trăm đến một triệu hình ảnh quả táo khác nhau, từ đó, gán nhãn, đào tạo cho AI hiểu về một "quả táo" sẽ có các pixel tương đồng ra sao, hình hài thế nào [9]. Mô hình học hỏi hình ảnh của AI có thể được biểu đạt trong minh họa sau:

Hình 5. Diễn tiến quá trình học tập về một vật thể của trí tuệ nhân tạo



Cũng bởi điều này, AI tạo sinh sẽ gặp khó khăn trong việc tái tạo một số chi tiết hình ảnh khó xử lý như bàn tay, bàn chân và chữ viết trong hình ảnh. Do vậy, phương án đầu tiên là quan sát kỹ các chi tiết thị giác bất thường, bao gồm răng, tay và nếp nhăn, làn da và dáng đi của một người. Những chi tiết nhỏ như tần suất chớp mắt, mụn, lỗ chân lông, xương hàm của người, biển hiệu, logo trên đường phố, nguồn sáng và hướng bóng đổ có thể không hợp lý, không tuân theo quy luật vật lý. Một số chi tiết càng nhỏ, AI tái tạo càng thiếu tự nhiên, như khuôn mặt đám đông, mồ hôi, tóc.

Đối với âm thanh, AI đôi lúc sẽ tạo ra khẩu hình không khớp âm lời nói, giọng điệu thiếu cảm xúc, quá trôi chảy. Một số mô hình như NotebookLM cố gắng tạo thêm các âm thanh ậm ừ, ngập ngừng như người thật, do vậy cần xét đến cả yếu tố tiếng động hiện trường, tiếng vang,… theo bối cảnh thực tế.

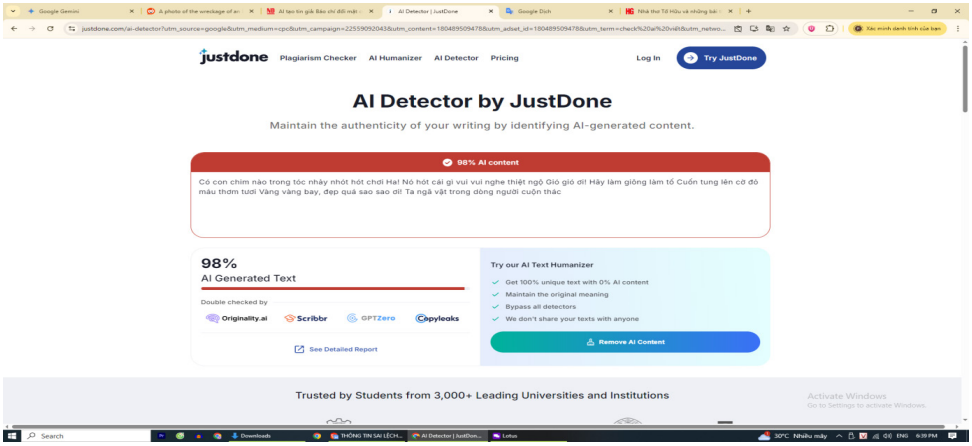
Đối với văn bản, AI có thể thường

xuyên viết hoa mà không có lý do, sử dụng các từ đệm khiến văn bản dài dòng mà không bổ sung nhiều ý nghĩa, cấu trúc cứng nhắc. AI đôi khi cũng trích nguồn, nhưng cần phải đọc lại nguồn để xem trang web có thật sự chứa đựng nội dung đó hay không, hoặc đơn giản là có…tồn tại hay không.

3.2. Phát hiện dựa trên AI và học máy

Dùng AI để phát hiện AI - đây là một ý tưởng không mới, song từ thực tế cho thấy, hiện chưa có công cụ thật sự hiệu quả để phát hiện một tin giả, bình luận giả tạo bởi AI. Ví dụ, như dùng công cụ Justdone để phát hiện yếu tố AI cho một đoạn thơ của nhà thơ Tố Hữu, AI này cho ra kết quả tới… 98%.

Hình 6. Công cụ phát hiện AI chưa thực sự hiệu quả



Trên thực tế, các thuật toán AI tỏ ra hiệu quả hơn để xác định các mẫu và dị vật trong hình ảnh và video để có thể chỉ ra tin giả [10]. Hơn nữa, AI có thể được sử dụng để theo dõi sự lan truyền của tin giả và xác định nguồn gốc của chúng nhằm ngăn chặn và giảm thiểu tốc độ lan truyền. Một số mô hình AI có thể kể đến để xác định yếu tố AI gồm:

- Tìm kiếm ngược hình ảnh (Reverse Image Search): *Google Images, TinEye, Yandex, PimEyes.*
- Phân tích Deepfake: *Intel FakeCatcher (Phân tích tín hiệu sinh học); McAfee Deepfake Detector (Sử dụng AI để xác định điểm bất thường; Deepware Scanner (Quét video để tìm dấu hiệu deepfake); Hive Moderation (Cung cấp API phát hiện hình ảnh/video AI).*

Tuy vậy, các mô hình được đào tạo trước thường hoạt động kém hiệu quả khi được thử nghiệm trên các loại deepfake mới mà không được tinh chỉnh (điểm AUC giảm xuống 67-71%) [1]. cho thấy một "khoảng cách tổng quát hóa" [12] trong phát hiện AI.

Những năm gần đây, các chuyên gia truyền thông, nhà báo cũng tận dụng AI như một "quân sư" giúp xác minh thông tin, xác định nguồn, kiểm tra chéo thông tin và cung cấp thêm ngữ cảnh để tăng tốc quá trình kiểm tra thực tế và cải thiện độ chính xác [11]. Trong khi các nhà báo giỏi trong việc hiểu ngữ cảnh và diễn giải sắc thái, thì AI vượt trội về quy mô và nhận dạng mẫu. Việc kết hợp những điểm mạnh này có thể tạo ra một hệ thống phát hiện mạnh mẽ hơn, khắc phục những hạn chế riêng lẻ. Như vậy, AI thay thế nỗ lực của con người mà tăng cường khả năng.

3.3. Phát hiện thông qua kiểm tra thực tế cộng đồng (Crowdsourcing)

Sức mạnh của số đông (còn gọi là Wisdom of crowds) đã được chứng minh là một phương pháp hiệu quả để xác định thông tin sai lệch; các nhóm hợp tác chính xác hơn các cá nhân trong việc phát hiện các bài đăng sai sự thật [12]. Ví dụ, có thể sử dụng Community Notes của X - một công cụ tận dụng phán đoán tập thể của người dùng để xác định và gắn cờ nội dung gây hiểu lầm [5].

Như vậy, có thể áp dụng phương pháp crowdsourcing (huy động sức mạnh đám đông) để phân tán gánh nặng xác minh. Chúng ta có thể đơn giản là đăng một nghi vấn lên trên một diễn đàn mạng xã hội và cùng phân tích với cộng đồng.

3.4. Nhận diện dựa vào thủy vân số

Thủy vân kỹ thuật số (hay Digital watermarking) là kỹ thuật nhúng các chữ ký không thể nhận biết nhưng có thể phát hiện được vào hình ảnh AI. Một số nhà phát triển AI (như Google DeepMind với SynthID) đã bắt đầu tích hợp các thủy vân số để nhận diện hình ảnh AI của họ. Ngoài

ra, người tạo hình ảnh (dù là con người hay AI) cũng có thể tự thêm thủy vân của họ (ví dụ: logo, văn bản "AI generated", tên tác giả) vào hình ảnh để bảo vệ bản quyền hoặc xác định nguồn gốc [13].

4. Kết luận và khuyến nghị

4.1. Kết luận

Nhìn chung, trong bối cảnh AI là trọng tâm của Cách mạng công nghiệp 4.0, AI tạo sinh sẽ phát triển không ngừng và ngày một tinh vi, qua đó, cũng khiến cuộc chiến với tin giả leo lên một bước thang mới. Bằng cách áp dụng một chiến lược toàn diện và thích ứng, tập trung vào cả vào đổi mới kỹ thuật, xây dựng chính sách và giáo dục người dùng, xã hội có thể giảm thiểu tác động tiêu cực của thông tin sai lệch và tin giả do AI tạo ra, bảo vệ tính toàn vẹn của thông tin và duy trì niềm tin trong không gian số.

Tại Hội thảo khoa học "*Sức mạnh không giới hạn và những thách thức khó dự báo của AI - Tác động và ứng phó chính sách*" ngày 15/9/2025, tại Hà Nội, Bộ Khoa học và Công nghệ cho biết đến cuối năm 2025, Việt Nam sẽ công bố bản cập nhật Chiến lược AI quốc gia cùng với Luật AI đầu tiên. Đây cũng là cơ sở để đội ngũ báo chí - truyền thông trong nước thúc đẩy các chiến lược về AI, nghiên cứu ứng dụng AI để thích ứng với thói quen tiêu thụ mới của khán giả, kiểm soát được các thách thức phức tạp của AI trên môi trường truyền thông.

4.2. Một số đề xuất và khuyến nghị nhằm giảm thiểu tác động của tin giả, tin sai sự thật từ nội dung đa phương tiện do AI tạo

4.2.1. Xây dựng khung pháp lý toàn diện khi đào tạo AI

Việc xây dựng khung pháp lý, thiết lập các tính năng an toàn cho mô hình AI ngay từ giai đoạn phát triển là yếu tố tiên quyết. Bên cạnh đó, trong bối cảnh các mô hình AI phát triển đa dạng, nhanh chóng, đã trở thành lăng kính phản ánh nhiều vấn đề; mang yếu tố tham khảo - hỗ trợ kiến thức, tin tức, học thuật cho nhiều đối tượng từ học sinh - sinh viên tới đội ngũ đào tạo, chuyên trách, rất cần chú trọng đến nhiệm vụ xây dựng các công cụ kiểm chứng nội dung AI. Đây là bài toán cần kết hợp giữa trách nhiệm ban đầu của nhà phát hành, với các công cụ kỹ thuật cao và các chương trình đào tạo, nâng cao năng lực phản biện cho người dùng. Khi các nội dung ngày một hoàn thiện ở bình diện siêu thực, các công cụ kiểm chứng cần được đồng bộ song song cùng quá trình phát triển các mô hình AI, đảm bảo các yếu tố từ phân tích - tìm kiếm - đánh giá bằng chứng là xác đáng, có cơ sở, có khả năng truy xuất nguồn gốc nội dung.

Từ đây, cũng đòi hỏi sự tham gia rộng rãi của liên minh quốc tế, chính phủ, các nhà khoa học xã hội, chuyên gia an ninh mạng, và cả người dùng thực tế. Các yếu tố cần cân nhắc bao gồm: nguồn dữ liệu đào tạo (độ sạch, độ đầy đủ, độ chính xác, số lượng, tính đa dạng (tránh thiên vị thông tin), tài nguyên tính toán, khả năng bảo mật dữ liệu, quy tắc đạo đức, quy trình quản lý,…). Xây dựng khung pháp lý cũng chỉ rõ trách nhiệm của các nhà phát triển, các nền tảng AI và người dùng khi sử dụng công nghệ này.

4.2.2. Nâng cao nhận thức và giáo dục về truyền thông trong bối cảnh AI

Do sự tinh vi ngày càng tăng của nội dung do AI tạo ra, việc chỉ dựa vào phát hiện kỹ thuật là không đủ. Giáo dục về truyền thông và truyền thông số, giáo dục về đạo đức công dân khi sử dụng AI là yếu tố then chốt để xây dựng khả năng cảnh giác, kỹ năng diễn giải, suy luận của xã hội đối với thông tin sai lệch và tin giả tạo bởi trí tuệ nhân tạo.

Sinh viên nói chung và sinh viên ngành Truyền thông đa phương tiện nói riêng cần sớm được tiếp cận với AI trong thực tế học tập, công việc để nắm bắt vấn đề hai mặt của hoạt động ứng dụng AI, không để sự hỗ trợ của AI làm xoá nhoà đi mục đích rèn luyện trí tuệ và giá trị đạo đức nghề nghiệp. Bên cạnh đó, sinh viên cũng cần nắm rõ vai trò trách nhiệm pháp lý, bản quyền và trách nhiệm xã hội trong quá trình sử dụng trí tuệ nhân tạo để giao tiếp, học tập, chia sẻ thông tin. Từ đó, có thể phát huy những thế mạnh riêng và kết hợp trải nghiệm, trái tim không thể thay thế của con người với sự bền bỉ, tốc độ và logic của trí tuệ nhân tạo.

4.2.3. Hợp tác đa bên

Sự hợp tác chặt chẽ giữa các nhà nghiên cứu AI, các nhà phát triển công nghệ, các nhà hoạch định chính sách, các tổ chức, các nhà giáo dục và cộng đồng người dùng sẽ tạo ra hệ sinh thái AI lành mạnh. Dẫu chuyển đổi công nghệ ra sao, thì con người vẫn là yếu tố mắt xích quan trọng nhất trong việc ngăn chặn tin giả, tin sai lệch tạo bởi AI, từ đó, giải quyết các thách thức đạo đức và pháp lý bao gồm thiên vị thuật toán, lo ngại về quyền riêng tư và bản chất "hộp đen" khi AI ra quyết định.

Danh mục tài liệu tham khảo

[1] Mỹ Linh and Hương Giang, "Khi AI tạo tin giả: Cuộc chiến vì sự thật của báo chí hiện đại" (video), 2025. [Online]. Available: https://nhandan.vn/video-khi-ai-tao-tin-gia-cuoc-chien-vi-su-that-cua-bao-chi-hien-dai-post887172.html. [Accessed: Jul. 2025].

[2] Trung tâm Truyền thông Khoa học và Công nghệ, "Việt Nam sắp công bố bản cập nhật Chiến lược AI quốc gia và Luật AI đầu tiên," 2025. [Online]. Available: https://mst.gov.vn/viet-nam-sap-cong-bo-ban-cap-nhat-chien-luoc-ai-quoc-gia-va-luat-ai-dau-tien-197250915155221269.htm. [Accessed: Sep. 15, 2025].

[3] R. Vinay, G. Spitale, N. Biller-Andorno, and F. Germani, "Emotional prompting amplifies disinformation generation in AI large language models," *Frontiers in Artificial Intelligence*, vol. 8, 2025.

[4] J. Gao, Y. Li, and T. Liu, "Generative artificial intelligence: a historical perspective," *National Science Review*, vol. 12, no. 5, 2024. [Online]. Available: https://doi.org/10.1093/nsr/nwaf050. [Accessed: Jun. 2025].

[5] J. G. Botha and H. Pieterse, "Fake news and deepfakes: A dangerous threat for 21st century information security," 2020. [Online]. Available: https://researchspace.csir.co.za/items/ace71ce8-6274-4bea-a393-7e06a68ab8e7. [Accessed: Jun. 2025].

[6] L. García-Faroldi, L. Teruel, and S. Blanco, "Unmasking AI's Role in the Age of Disinformation: Friend or Foe?," *Journalism and Media*, vol. 6, no. 1, 2025. [Online]. Available: https://www.mdpi.com/2673-5172/6/1/19.

[7] M. Appel and F. Prietzel, "The detection of political deepfakes," *Journal of Computer-Mediated Communication*, vol. 27, no. 4, 2022. [Online]. Available: https://doi.org/10.1093/jcmc/zmac008. [Accessed: Jun. 2025].

[8] R. Shetty, "Top 10 Terrifying Deepfake Examples," Arya.ai, 2025. [Online]. Available: https://arya.ai/blog/top-deepfake-incidents.

[9] A. R. Chow and B. Perrigo, "Google's Veo 3 Can Make Deepfakes of Conflict, Riots, More," *TIME*, 2025. [Online]. Available: https://time.com/7290050/veo-3-google-misinformation-deepfake/.

[10] University of Toronto Libraries, "How Generative AI Models Work," 2025. [Online]. Available: https://guides.library.utoronto.ca/c.php?g=735513&p=5297039. [Accessed: Jun. 2025].

[11] T. Lrx, "The generalisability gap: Evaluating deepfake detectors across domains," *Quasilinear Musings*, 2025. [Online]. Available: https://www.timlrx.com/blog/how-generalisable-are-deepfake-detectors.

[12] R. LaForme, "How Sora, OpenAI's new text-to-video tool, could harm journalism and society," *Poynter*, 2024. [Online]. Available: https://www.poynter.org/tech-tools/2024/openai-sora-effects-journalism-democracy-misinformation/.

[13] A. Diel et al., "Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers," *Computers in Human Behavior Reports*, vol. 16, 100538, 2024.

[14] H. Luo, L. Li, and J. Li, "Digital Watermarking Technology for AI-Generated Images: A Survey," *Mathematics*, vol. 13, no. 4, 651, 2025. [Online]. Available: https://doi.org/10.3390/math13040651. [Accessed: May 2025].

[15] T. B. Carl and J. D. W. Jevin, *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. New York: Random House Publishing Group, 2020.

[16] A. Ng, "Artificial intelligence is the new electricity," Stanford University, 2017. [Online]. Available: www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity.

[17] C. Jia, "Collaboration, crowdsourcing, and misinformation," *Internet Democracy Initiative*, Northeastern University, 2024. [Online]. Available: https://idi.provost.northeastern.edu/2024/09/13/collaboration-crowdsourcing-and-misinformation/.

[18] HyperVerge, "Top 10 Examples of Deepfake Across The Internet," 2025. [Online]. Available: https://hyperverge.co/blog/examples-of-deepfakes/. [Accessed: Jun. 2025].

[19] L. Lusquino Filho and A. Rocha, "The rise of synthetic realities: Impact, advancements, and ethical considerations," *IEEE Systems, Man, and Cybernetics Magazine*, 2024. [Online]. Available: https://www.ieeesmc.org/wp-content/uploads/2024/03/FeatureArticle_03_2024.pdf. [Accessed: Jun. 2025].