

NĂNG LỰC NHẬN DIỆN DEEFAKE TRONG GIÁO DỤC CÔNG DÂN SỐ CHO SINH VIÊN ĐẠI HỌC MỞ: KHUNG TIẾP CẬN TỪ TƯ DUY PHÊ PHÁN VÀ ĐẠO ĐỨC AI

Nguyễn Thu Hà¹, Tạ Ngọc Như Quỳnh¹, Đỗ Đoàn Nhật Lệ¹

Email: ntha13@hou.edu.vn. ORCID: 0009-0001-1103-4850

Ngày tòa soạn nhận được bài báo: 06/07/2026

Ngày hoàn thành đánh giá: 05/10/2026

Ngày bài báo được duyệt đăng: 19/10/2026

DOI: 10.59266/houjs.2026.1494

Tóm tắt: Bài viết phân tích nhu cầu phát triển năng lực nhận diện deepfake (nội dung giả mạo tổng hợp bằng trí tuệ nhân tạo) trong giáo dục công dân số (digital citizenship) cho sinh viên Trường Đại học Mở Hà Nội trong bối cảnh trí tuệ nhân tạo tạo sinh (Generative Artificial Intelligence - GenAI) đang làm thay đổi cách người học tiếp cận, đánh giá và chia sẻ thông tin. Trên cơ sở tổng hợp tài liệu và phân tích chính sách, nghiên cứu đề xuất khung năng lực khái niệm (chưa qua kiểm định thực nghiệm trên người học) gồm năm thành tố: (1) Hiểu biết về truyền thông tổng hợp (synthetic media); (2) Kiểm chứng nguồn và bối cảnh qua kỹ năng đọc ngang (lateral reading); (3) Tư duy phê phán (critical thinking); (4) Đạo đức trí tuệ nhân tạo (AI ethics); và (5) Thực hành công dân số có trách nhiệm. Bài viết gợi ý thiết kế mô-đun 8 tiết tích hợp vào học phần kỹ năng số hoặc công dân số cho sinh viên Trường Đại học Mở Hà Nội, kèm theo bộ tiêu chí đánh giá (rubric 4 mức độ). Nghiên cứu đóng góp hướng tiếp cận giáo dục phòng ngừa rủi ro truyền thông tổng hợp, đặt trọng tâm vào năng lực phán đoán, liêm chính học thuật và trách nhiệm công dân số của người học.

Từ khóa: công dân số, deepfake, đạo đức trí tuệ nhân tạo, sinh viên đại học mở, tư duy phê phán

I. Đặt vấn đề

Trí tuệ nhân tạo tạo sinh (Generative Artificial Intelligence - GenAI) đang đưa giáo dục đại học vào một giai đoạn mới, nơi văn bản, hình ảnh, âm thanh và video có thể được tạo ra hoặc biến đổi với tốc độ rất nhanh. Trong dòng chảy đó, công nghệ deepfake - được hiểu là dạng nội dung giả mạo tổng hợp do trí

tuệ nhân tạo (AI) tạo ra nhằm mô phỏng hoặc thay đổi hình ảnh, giọng nói, khuôn mặt hay hành vi của một cá nhân - đã trở thành một thách thức trực tiếp đối với môi trường học tập số. Khi công cụ tạo nội dung ngày càng dễ sử dụng, ranh giới giữa nội dung thật, nội dung chỉnh sửa và nội dung tổng hợp trở nên mong manh hơn.

¹ Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

Đối với sinh viên Trường Đại học Mở Hà Nội và mô hình đại học mở nói chung, vấn đề này càng có ý nghĩa thực tiễn. Mô hình đại học mở nhấn mạnh khả năng tiếp cận giáo dục, học tập linh hoạt, học tập trực tuyến, tự học và khai thác tài nguyên giáo dục mở. Những đặc trưng đó giúp mở rộng cơ hội học tập nhưng cũng khiến sinh viên tiếp xúc thường xuyên hơn với video bài giảng, nền tảng mạng xã hội, trang web thông tin và các diễn đàn trực tuyến ngoài lớp học truyền thống. Trong thực tế, các rủi ro từ truyền thông tổng hợp (synthetic media) đã xuất hiện dưới dạng video mạo danh thông báo học phí/lich thi, hình ảnh cắt ghép của cán bộ/giảng viên, hoặc việc sinh viên lạm dụng AI tạo lập nội dung giả mạo trong sản phẩm học tập. Nếu thiếu năng lực đánh giá nguồn tin và nhận diện nội dung tổng hợp, sinh viên có thể bị tác động bởi thông tin sai lệch, bị cuốn vào hành vi chia sẻ thiếu kiểm chứng hoặc vi phạm liên chính học thuật.

Các nghiên cứu quốc tế cho thấy deepfake trong giáo dục đại học vừa tạo rủi ro vừa có khả năng ứng dụng tích cực. Roe và cộng sự (2024) nhận định nghiên cứu về deepfake trong giáo dục đại học còn hạn chế, trong khi các rủi ro về danh tính, niềm tin và gian lận học thuật đang gia tăng. Hwang và cộng sự (2021) cho thấy giáo dục truyền thông có thể giảm tác động thuyết phục và ý định chia sẻ thông tin sai lệch có sử dụng deepfake.

Các nghiên cứu quốc tế cho thấy deepfake trong giáo dục đại học vừa tạo rủi ro vừa có khả năng ứng dụng tích cực (Roe và cộng sự, 2024; Westerlund, 2019). Hwang và cộng sự (2021) chỉ ra rằng giáo dục truyền thông có thể giảm tác động thuyết phục và ý định chia sẻ thông tin sai

lệch do AI tạo ra. Tại Việt Nam, Nguyễn và cộng sự (2025) đã đề xuất khung năng lực số cho sinh viên từ xa trực tuyến tại Trường Đại học Mở Hà Nội, trong khi Đỗ và Lại (2026) ghi nhận những hạn chế về năng lực tự học và kiểm chứng thông tin của sinh viên trong môi trường số. Tuy nhiên, các công trình này chưa xây dựng một bộ tiêu chí vận dụng chuyên biệt về nhận diện deepfake và đạo đức AI cho người học trong môi trường đại học mở.

Từ khoảng trống đó, bài viết tập trung giải quyết hai câu hỏi nghiên cứu cụ thể:

RQ1: *Những thành tố năng lực cốt lõi nào cần thiết để sinh viên nhận diện, kiểm chứng và ứng xử có trách nhiệm trước nội dung deepfake trong bối cảnh giáo dục công dân số?*

RQ2: *Các thành tố năng lực ấy có thể được chuyển hóa thành bộ tiêu chí đánh giá (rubric) và mô-đun học tập tích hợp cho sinh viên Trường Đại học Mở Hà Nội như thế nào?*

Bài viết kế thừa Khung năng lực số cho công dân DigComp 2.2 (Vuorikari và cộng sự, 2022), các nguyên tắc công dân số (Ribble, 2015), tư duy phê phán và đạo đức AI để đề xuất một khung khái niệm tích hợp theo bối cảnh. Bài viết xác định rõ đây là nghiên cứu tổng hợp lý thuyết, không xem việc không sử dụng khảo sát thực nghiệm là điểm mới của nghiên cứu.

II. Cơ sở lý thuyết

2.1. Deepfake và truyền thông tổng hợp trong giáo dục đại học

Deepfake là dạng truyền thông tổng hợp (synthetic media) do công nghệ AI tạo ra thông qua các thuật toán học sâu, trong đó hình ảnh, âm thanh hoặc

video của một người bị mô phỏng, thay thế hoặc biến đổi. Karnouskos (2020) cho rằng sự phát triển của deepfake đặt ra các vấn đề xã hội đáng chú ý liên quan đến niềm tin, xác thực thông tin, danh tiếng cá nhân và quản trị truyền thông. Trong giáo dục đại học, deepfake có thể xuất hiện dưới dạng giả mạo phát biểu của giảng viên, tạo video/giọng nói của người khác khi chưa được đồng ý, phát tán nội dung sai về nhà trường, hoặc sử dụng hình ảnh tổng hợp trong bài nộp mà không công bố nguồn gốc.

Tuy nhiên, Roe và cộng sự (2024) nhận định truyền thông tổng hợp cũng có thể hỗ trợ tạo học liệu minh họa, mô phỏng tình huống và tăng tính tương tác nếu được kiểm soát tốt. Vấn đề trọng tâm là xây dựng năng lực để sinh viên biết khi nào và bằng cách nào nội dung tổng hợp được sử dụng minh bạch và tuân thủ liên chính học thuật.

2.2. Công dân số và năng lực số của sinh viên đại học mở

Lý thuyết công dân số (digital citizenship) nhấn mạnh trách nhiệm, chuẩn mực hành vi và năng lực tham gia an toàn trong môi trường số (Ribble, 2015). Đối với sinh viên đại học mở, công dân số đòi hỏi khả năng tự học tự chủ, kiểm chứng thông tin, tôn trọng quyền riêng tư và ứng xử văn minh trong cộng đồng học thuật trực tuyến. Nghiên cứu của Wong và Abas (2023) trên Tạp chí Giáo dục Mở và Từ xa Châu Á (*Asian Journal of Open and Distance Learning* - AJODL) cũng khẳng định rằng sinh viên học tập linh hoạt cần được trang bị năng lực thích ứng số và năng lực AI để chủ động phòng ngừa rủi ro thao túng thông tin.

Tại Việt Nam, Quyết định số 146/QĐ-TTg ngày 28/01/2022 của Thủ tướng Chính phủ đã phê duyệt Đề án nâng cao nhận thức và phổ cập kỹ năng số quốc gia, tạo hành lang chính sách cho việc phát triển năng lực số trong các cơ sở đào tạo. Khung năng lực số đặc thù cho sinh viên đại học từ xa trực tuyến tại Trường Đại học Mở Hà Nội do Nguyễn và cộng sự (2025) đề xuất và đánh giá năng lực tự học của Đỗ và Lại (2026) công bố trên Tạp chí Khoa học Trường Đại học Mở Hà Nội đều nhấn mạnh vai trò của môi trường học tập số và yêu cầu kiểm chứng thông tin tự định hướng của người học.

Bên cạnh đó, Khung năng lực số cho công dân DigComp 2.2 (Vuorikari và cộng sự, 2022) cung cấp các chỉ báo cập nhật về giao tiếp, sáng tạo nội dung, an toàn số và ứng dụng AI. Đây là cơ sở quan trọng để định hình các hành vi công dân số trước rủi ro truyền thông tổng hợp.

2.3. Tư duy phê phán và năng lực kiểm chứng thông tin

Tư duy phê phán (critical thinking) là năng lực suy nghĩ hợp lý và phản tư nhằm quyết định điều nên tin và nên làm (Ennis, 1987). Paul và Elder (2008) nhấn mạnh các chuẩn mực trí tuệ như độ chính xác, tính logic, sự rõ ràng và tính công bằng trong đánh giá lập luận. Trước rủi ro deepfake, tư duy phê phán giúp sinh viên không dừng lại ở cảm giác trực quan bề ngoài mà phân tích sâu về bối cảnh, động cơ và bằng chứng hỗ trợ.

Để kiểm chứng hiệu quả, Wineburg và McGrew (2019) đề xuất kỹ năng “đọc ngang” (lateral reading) - thao tác nhanh chóng rời khỏi trang web đang xem để tra

cứu, đối chiếu nguồn tin từ các trang độc lập khác. Kỹ năng này đóng vai trò then chốt vì các dấu hiệu thị giác bề ngoài của deepfake ngày càng khó phát hiện bằng mắt thường.

2.4. Đạo đức AI và trách nhiệm học thuật

Đạo đức AI (AI ethics) bao gồm các nguyên tắc cốt lõi như minh bạch, công bằng, trách nhiệm giải trình, quyền riêng tư và không gây hại (Jobin và cộng sự, 2019). Hướng dẫn của UNESCO về AI tạo sinh trong giáo dục và nghiên cứu (Miao và Holmes, 2023) nhấn mạnh nguyên tắc lấy con người làm trung tâm và tính minh bạch trong việc công bố sử dụng công cụ AI. Đối với sinh viên, đạo đức AI gắn liền với liêm chính học thuật: công khai mức độ can thiệp của AI, không giả mạo danh tính, tôn trọng quyền hình ảnh/giọng nói của bạn học và giảng viên, và không tạo lập nội dung gây tổn hại cho cộng đồng.

III. Phương pháp nghiên cứu

Bài viết sử dụng phương pháp nghiên cứu lý thuyết, không tiến hành khảo sát, phỏng vấn hay thực nghiệm trên người học. Quy trình nghiên cứu gồm ba bước:

Thứ nhất, nghiên cứu thực hiện tổng quan tài liệu có định hướng đối với 28 nguồn tài liệu khoa học và văn bản chính sách. Các tài liệu được tìm kiếm trên Google Scholar, VJOL, trang web của Tạp chí Khoa học Trường Đại học Mở Hà Nội (JSHOU) và Tạp chí AJODL. Tiêu chí chọn tài liệu dựa trên tính liên quan đến rủi ro của truyền thông tổng

hợp, năng lực công dân số, tư duy phê phán, đạo đức AI và sự phù hợp với bối cảnh đại học mở.

Thứ hai, nghiên cứu phân tích Khung năng lực số cho công dân DigComp 2.2 (Vuorikari và cộng sự, 2022), Hướng dẫn của UNESCO (Miao và Holmes, 2023), Quyết định 146/QĐ-TTg của Thủ tướng Chính phủ và các nghiên cứu về năng lực số sinh viên Trường Đại học Mở Hà Nội (Nguyễn và cộng sự, 2025) để xác định các yêu cầu hành vi chuẩn.

Thứ ba, nghiên cứu thực hiện tổng hợp lý thuyết để xây dựng ma trận đối chiếu và đề xuất khung năng lực 5 thành tố. Ranh giới giữa hai thành tố dễ trùng lặp được phân định như sau: “*Kiểm chứng nguồn và bối cảnh*” là các thao tác kỹ thuật ngoại cảnh (đọc ngang, tra cứu vết số, đối chiếu trang web chính thức); trong khi “*Tư duy phê phán*” là quá trình phân đoán logic nội tại (đánh giá sức nặng bằng chứng, phát hiện bias cảm xúc và suy luận hệ quả).

Ghi chú: Khung năng lực này mang tính chất một khung khái niệm đề xuất, chưa được thẩm định qua kỹ thuật Delphi hay thử nghiệm thực nghiệm, phục vụ định hướng thiết kế nội dung giáo dục.

IV. Kết quả nghiên cứu và thảo luận

4.1. Quy trình hình thành và Ma trận đối chiếu khung năng lực

Năm thành tố của khung năng lực được rút ra từ việc đối chiếu giữa lý thuyết nền tảng và yêu cầu thực tiễn của sinh viên đại học mở, trình bày cụ thể tại Bảng 1.

Bảng 1. Ma trận đối chiếu hình thành khung năng lực nhận diện deepfake

Thành tố năng lực	Nguồn lý thuyết & Văn bản tham chiếu	Đặc thù sinh viên Trường Đại học Mở	Phân định chức năng
1. Hiểu biết về truyền thông tổng hợp	Karnouskos (2020); Roe và cộng sự (2024); DigComp 2.2 (Vuorikari và cộng sự, 2022)	Thường xuyên tiếp xúc học liệu đa phương tiện, bài giảng video trực tuyến.	Hiểu bản chất công nghệ; phân biệt nội dung gốc, nội dung chỉnh sửa và nội dung do AI tạo ra.
2. Kiểm chứng nguồn và bối cảnh	Wineburg và McGrew (2019); Nguyễn và cộng sự (2025); Wong và Abas (2023)	Tự khai thác tài nguyên ngoài giáo trình; đối mặt với nguy cơ mạo danh thông báo nhà trường.	Thao tác ngoại cảnh: Đọc ngang, kiểm tra dấu vết số, đối chiếu trang web chính thức.
3. Tư duy phê phán	Ennis (1987); Paul và Elder (2008); Đỗ và Lại (2026)	Môi trường tự học đòi hỏi khả năng phán đoán độ tin cậy của thông tin trước khi chia sẻ.	Phán đoán nội tại: Đánh giá logic lập luận, phát hiện bấy cảm xúc, phân tích động cơ truyền thông.
4. Đạo đức AI	Jobin và cộng sự (2019); Miao và Holmes (2023)	Tương tác trực tuyến nhiều, áp lực làm bài tập số, nguy cơ vi phạm liên chính học thuật.	Tôn trọng quyền hình ảnh, minh bạch phạm vi sử dụng AI, tuân thủ liên chính học thuật.
5. Thực hành công dân số có trách nhiệm	Ribble (2015); Quyết định 146/QĐ-TTg (Thủ tướng Chính phủ, 2022)	Tham gia các diễn đàn, nhóm mạng xã hội của lớp và trường.	Báo cáo nội dung độc hại/giả mạo, chủ động xây dựng môi trường học tập an toàn.

Nguồn: Tác giả tổng hợp (2026)

4.2. Khung năng lực nhận diện deepfake cho sinh viên đại học mở

Khung năng lực đề xuất gồm 5 thành tố, đi kèm các biểu hiện hành vi và ví dụ vận dụng trong bối cảnh Trường Đại học Mở Hà Nội (Bảng 2).

Bảng 2. Khung năng lực nhận diện deepfake cho sinh Trường Đại học Mở

Thành tố năng lực	Biểu hiện cần đạt	Ví dụ trong môi trường Trường Đại học Mở
1. Hiểu biết về truyền thông tổng hợp	- Giải thích được khái niệm deepfake và truyền thông tổng hợp. - Nhận biết các dạng deepfake (khuôn mặt, giọng nói, video, văn bản). - Hiểu rủi ro mạo danh lẫn cơ hội ứng dụng của AI.	Nhận biết một video thông báo của nhà trường có dấu hiệu bất thường về khẩu hình hoặc âm thanh.
2. Kiểm chứng nguồn và bối cảnh	- Thực hành đọc ngang để kiểm tra thông tin trên các trang web độc lập. - Xác định thời điểm, nguồn phát tán và bằng chứng đối chiếu. - Tra cứu xác minh trên cổng thông tin chính thức của nhà trường.	Đối chiếu thông tin lịch thi nhận từ mạng xã hội với thông báo trên trang web chính thức.
3. Tư duy phê phán	- Phân tích tính logic, sự nhất quán và bằng chứng của nội dung. - Nhận diện các kỹ thuật thao túng cảm xúc (kích động, gây sợ hãi). - Đặt câu hỏi về mục đích và động cơ của bên phát tán.	Không vội chia sẻ video gây tranh cãi về giảng viên khi chưa đánh giá động cơ và nguồn độc lập.

Thành tố năng lực	Biểu hiện cần đạt	Ví dụ trong môi trường Trường Đại học Mở
4. Đạo đức AI	- Tôn trọng quyền riêng tư, danh tính và quyền hình ảnh cá nhân. - Tự nguyện công bố minh bạch mức độ sử dụng AI trong bài nộp. - Tuân thủ liên chính học thuật, không tạo nội dung giả mạo.	Khai báo rõ trong bài tập nhóm: “Hình ảnh minh họa do AI tạo ra theo gợi ý của nhóm”.
5. Thực hành công dân số có trách nhiệm	- Báo cáo nội dung giả mạo, lừa đảo cho ban quản lý LMS hoặc nhà trường. - Không lan truyền thông tin chưa được kiểm chứng. - Hỗ trợ bạn học nhận diện rủi ro và ứng xử an toàn trên mạng.	Báo cáo cho bộ phận hỗ trợ kỹ thuật khi phát hiện tài khoản mạo danh cán bộ lớp phát tán đường link lạ.

Nguồn: Tác giả đề xuất (2026)

4.3. Bộ tiêu chí đánh giá năng lực nhận diện deepfake (Rubric)

Để hỗ trợ công tác đo lường và định hướng giảng dạy, bài viết xây dựng Bộ tiêu chí đánh giá (Rubric) gồm 4 mức độ (Bảng 3).

Bảng 3. Rubric đánh giá năng lực nhận diện deepfake của sinh viên

Thành tố	Mức 1: Chưa đạt	Mức 2: Đạt một phần	Mức 3: Đạt	Mức 4: Thành thạo
1. Hiểu biết truyền thông tổng hợp	Không phân biệt được nội dung gốc và sản phẩm AI tổng hợp.	Nêu được khái niệm AI nhưng chưa chỉ ra các dạng deepfake cụ thể	Phân biệt rõ các dạng deepfake và nêu được rủi ro lẫn ứng dụng.	Giải thích sâu nguyên lý AI tạo sinh và dự báo rủi ro truyền thông mới
2. Kiểm chứng nguồn & bối cảnh	Tin tưởng cảm quan bề ngoài; chỉ đọc đọc trên trang web chứa thông tin	Biết nghi ngờ nhưng chỉ kiểm tra cảm quan bề mặt.	Thực hành thành thạo đọc ngang, đối chiếu được 2 nguồn độc lập	Sử dụng chuyên sâu các công cụ tra cứu vết số và xác minh nguồn gốc
3. Tư duy phê phán	Đánh giá dựa trên cảm xúc; vội vã chia sẻ thông tin giật gân.	Nhận ra thông tin bất thường nhưng chưa chỉ ra được lỗi logic.	Phân tích bài bản logic, bày cảm xúc và động cơ phát tán.	Đánh giá toàn diện tác động xã hội và các kỹ thuật thao túng tinh vi
4. Đạo đức AI	Tự ý cắt ghép ảnh/giọng nói người khác; giấu việc dùng AI.	Dùng AI tạo nội dung nhưng khai báo sơ sài hoặc chưa tự nguyện	Tôn trọng bản quyền, công khai minh bạch phạm vi sử dụng AI	Nêu cao liên chính học thuật, chủ động bảo vệ dữ liệu cá nhân
5. Thực hành công dân số	Thụ động; vô tình lan truyền tin giả hoặc nội dung deepfake.	Nhận biết tin giả nhưng không báo cáo cho cơ quan quản lý.	Chủ động báo cáo nội dung giả mạo, không chia sẻ tin chưa kiểm chứng	Hướng dẫn cộng đồng học tập ứng xử an toàn trước rủi ro AI.

Nguồn: Tác giả đề xuất (2026)

4.4. Đề xuất mô-đun đào tạo tích hợp 8 tiết

Bài viết đề xuất mô-đun tích hợp ngắn hạn mang tên “*Nhận diện Deepfake và Đạo đức AI trong Môi trường Số*”, thời lượng 8 tiết (4 tiết lý thuyết + 4 tiết thực hành), có thể lồng ghép vào học phần Kỹ năng số hoặc Công dân số cho sinh viên Trường Đại học Mở Hà Nội (Bảng 4).

Chuẩn đầu ra của mô-đun (CLOs):

- CLO1: Phân tích được bản chất, các dạng thức và rủi ro của deepfake trong học tập số.

- CLO2: Áp dụng kỹ năng đọc ngang để kiểm chứng nguồn và bối cảnh của thông tin nghi vấn.

- CLO3: Vận dụng tư duy phê phán để đánh giá bằng chứng và động cơ truyền thông.

- CLO4: Thực hành ứng xử tuân thủ đạo đức AI, Liêm chính học thuật và trách nhiệm công dân số.

Bảng 4. Cấu trúc mô-đun đào tạo 8 tiết đề xuất

Chủ đề / Thời lượng	CLOs	Nội dung trọng tâm	Hoạt động dạy và học	Phương pháp đánh giá
Chủ đề 1: Tổng quan Deepfake & AI Tạo sinh (2 tiết)	CLO1	- Khái niệm truyền thông tổng hợp và deepfake. - Các dạng thức và rủi ro trong giáo dục đại học.	- Bài giảng tương tác. - Phân tích mẫu video/âm thanh thật và do AI tạo ra.	Bài trắc nghiệm ngắn (Quiz) kiểm tra nhận diện dấu hiệu AI.
Chủ đề 2: Kỹ thuật Kiểm chứng Nguồn & Đọc ngang (2 tiết)	CLO2	- Kỹ năng đọc ngang (lateral reading). - Quy trình tra cứu vết số và trang web chính thức	- Thực hành cá nhân/nhóm trên máy tính. - Thực hiện kiểm chứng 2 tin tức nghi vấn trên mạng.	Hồ sơ kiểm chứng thông tin (Checklist đối chiếu nguồn).
Chủ đề 3: Tư duy Phê phán trước Nội dung AI (2 tiết)	CLO3	- Nhận diện kỹ thuật thao túng cảm xúc. - Đánh giá bằng chứng và logic lập luận.	- Thảo luận tình huống (Case study) về tin đồn mạo danh cán bộ/giảng viên.	Bản lập luận phản tư (Reflective writing 300 từ).
Chủ đề 4: Đạo đức AI & Liêm chính Học thuật (2 tiết)	CLO4	- Quyền hình ảnh và dữ liệu cá nhân. - Nguyên tắc công khai AI và liêm chính học thuật.	- Đóng vai xử lý tình huống vi phạm đạo đức AI. - Xây dựng Bảng quy tắc ứng xử cá nhân.	Đánh giá bảng quy tắc cá nhân và phần thể hiện đóng vai.

4.5. Thảo luận

So với khung năng lực số tổng quát của Nguyễn và cộng sự (2025), nghiên cứu này đi sâu giải quyết rủi ro chuyên biệt từ truyền thông tổng hợp do AI tạo ra. Kết quả cũng hỗ trợ các nhận định của Đỗ và Lại (2026) về tầm quan trọng của việc xây dựng tiêu chí đánh giá rõ ràng nhằm hỗ trợ năng lực tự học của sinh viên.

Đóng góp chính của bài viết là chuyển dịch trọng tâm từ giải pháp kỹ thuật thụ động sang nâng cao năng lực chủ động của con người: hiểu biết công nghệ, kiểm chứng thông tin, tư duy phê phán, đạo đức AI và trách nhiệm công dân số. Cách tiếp cận này nhất quán với các khuyến nghị quốc tế của Hwang và cộng sự (2021) cũng như Miao và Holmes (2023).

V. Kết luận và kiến nghị

Bài viết đã làm rõ cơ sở lý thuyết và đề xuất khung năng lực nhận diện deepfake 5 thành tố phù hợp với sinh viên đại học mở. Nghiên cứu đã chuyển hóa khung khái niệm này thành bộ tiêu chí đánh giá (rubric 4 mức độ) và mô-đun đào tạo 8 tiết có chuẩn đầu ra rõ ràng cho sinh viên Trường Đại học Mở Hà Nội.

Bài viết đề xuất các kiến nghị sau:

1. Trường Đại học Mở Hà Nội:

Xem xét tích hợp mô-đun 8 tiết vào chương trình kỹ năng số hoặc công dân số; ban hành quy định hướng dẫn liên môn chính học thuật về sử dụng AI tạo sinh.

2. **Giảng viên:** Bồi dưỡng kỹ năng cập nhật về AI và kiểm chứng thông tin để lồng ghép các bài tập tình huống thực tế vào giảng dạy.

3. Hướng nghiên cứu tiếp theo:

Do nghiên cứu mới dừng ở mức xây dựng khung khái niệm và đề xuất mô-đun sơ phạm, các nghiên cứu tương lai cần tiến hành tham vấn chuyên gia (Delphi) và triển khai thực nghiệm (đo lường trước - sau can thiệp) trên sinh viên để đánh giá hiệu quả thực tế của khung năng lực.

Tài liệu tham khảo

- Đỗ, T. H. N., và Lại, T. Y. N. (2026). Đánh giá năng lực tự học của sinh viên đại học trong môi trường số. *Tạp chí Khoa học Trường Đại học Mở Hà Nội*, (Số đặc biệt 11A), 45-56. <https://doi.org/10.59266/houjs.2026.1219>
- Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills* (pp. 9-26). W. H. Freeman. <https://archive.org/details/teachingthinking00hing>
- Hobbs, R. (2010). *Digital and media literacy: A plan of action*. The Aspen Institute. https://www.criticalthinking.org/files/Concepts_Tools.pdf
- Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, 24(3), 188-193. <https://doi.org/10.1089/cyber.2020.0174>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. <https://doi.org/10.1038/s42256-019-0088-2>
- Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, 1(3), 138-147. <https://doi.org/10.1109/TTS.2020.3001312>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing. <https://doi.org/10.54675/ASFR2115>
- Nguyễn, Q. A., Nguyễn, V. H., và Nguyễn, H. S. (2025). Phát triển khung năng lực số đặc thù cho sinh viên đại học từ xa trực tuyến: Bối cảnh tại Trường Đại học Mở Hà Nội. *Tạp chí Khoa học Trường Đại học Mở Hà Nội*, (134), 12-23. <https://doi.org/10.59266/houjs.2025.1088>
- Paul, R., & Elder, L. (2008). *The miniature guide to critical thinking: Concepts and tools* (5th ed.). Foundation for Critical Thinking. https://www.ubiquityuniversity.org/wp-content/uploads/2018/02/Concepts_-_Critical-Thinking-HandbookTools-Ubiquity-University.pdf
- Ribble, M. (2015). *Digital citizenship in schools: Nine elements all students should know* (3rd ed.). International Society for Technology in Education. [cite: 1, 3, 4, 6] https://archive.org/details/details/digitalcitizensh0000ribb_x4u3

- Roe, J., Perkins, M., & Furze, L. (2024). Deepfakes and higher education: A research agenda and scoping review of synthetic media. *Journal of University Teaching and Learning Practice*, 21(10), 1-22. <https://doi.org/10.53761/2y2np178>
- Thủ tướng Chính phủ. (2022). *Quyết định số 146/QĐ-TTg ngày 28/01/2022 phê duyệt Đề án “Nâng cao nhận thức, phổ cập kỹ năng và phát triển nguồn nhân lực chuyển đổi số quốc gia đến năm 2025, định hướng đến năm 2030”*. Công Thông tin điện tử Chính phủ. [cite: 3, 4, 6] <https://vanban.chinhphu.vn/?pageid=27160&docid=205276>
- Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2: The digital competence framework for citizens - With new examples of knowledge, skills and attitudes*. Publications Office of the European Union. <https://doi.org/10.2760/115376>
- Westerlund, M. (2019). The emergence of deepfakes. *Technology Innovation Management Review*, 9(11), 39-52. <https://doi.org/10.22215/timreview/1282>
- Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, 121(11), 1-40. https://stacks.stanford.edu/file/druid:yk133ht8603/Wineburg%20McGrew_Lateral%20Reading%20and%20the%20Nature%20of%20Expertise.pdf
- Wong, L., & Abas, Z. W. (2023). Building digital resilience and AI literacy in open and distance learning: A regional perspective. *Asian Journal of Open and Distance Learning*, 11(2), 34-49. <https://ajodl.oum.edu.my/>

DEEFAKE LITERACY IN DIGITAL CITIZENSHIP EDUCATION FOR OPEN UNIVERSITY STUDENTS: A FRAMEWORK BASED ON CRITICAL THINKING AND AI ETHICS

Nguyen Thu Ha¹, Ta Ngoc Nhu Quynh¹, Do Doan Nhat Le¹

Abstract: *This paper examines the need to develop deepfake literacy within digital citizenship education for Hanoi Open University students in the context of generative artificial intelligence (GenAI). Based on a directed literature review and policy document analysis, the study proposes a conceptual competency framework (not yet empirically tested on learners) comprising five components: (1) Understanding synthetic media; (2) Verifying sources and contexts via lateral reading; (3) Exercising critical thinking; (4) Applying AI ethics; and (5) Practising responsible digital citizenship. The paper suggests an 8-period learning module with a 4-level rubric assessment grid designed for integration into digital skills or digital citizenship courses for Hanoi Open University students. The study contributes an educationally oriented approach to mitigating synthetic media risks by shifting attention from purely technical detection to critical judgment, academic integrity, and ethical digital citizenship.*

Keywords: *AI ethics, critical thinking, deepfake, digital citizenship, open university students*

¹ Hanoi, Open University, Hanoi, Vietnam