ĐÁNH GIÁ CƠ CHẾ ỨNG PHÓ VỚI TIN GIẢ DO DEEPFAKE TRÊN MẠNG XÃ HỘI TẠI VIỆT NAM VÀ ĐỀ XUẤT CÁC KHUYẾN NGHỊ LIÊN QUAN

VŨ THỊ MINH TÂM
Trường Đại học Hạ Long

Nhận bài ngày 11/11/2025. Sửa chữa xong 31/12/2025. Duyệt đăng 13/01/2026.

Abstract

The rapid proliferation of artificial intelligence technologies, particularly deepfakes, has posed significant challenges to information security on social media platforms in Vietnam. This article evaluates the current status and effectiveness of existing response mechanisms, including early-warning systems, identification and verification processes, and social intervention measures. The findings reveal systemic vulnerabilities, such as engagement-driven algorithms that undermine warning efforts, reliance on foreign detection technologies lacking local specificity, and legal enforcement difficulties in addressing cross-border violations. Based on this analysis, the study proposes a comprehensive framework of solutions, including the localization of AI-based detection technologies, the establishment of a multi-layered verification ecosystem, the strengthening of legal sanctions, and the development of digital resilience through community education. The article contributes both theoretical and practical insights to digital media governance and the protection of cybersecurity in Vietnam.

Keywords: *Deepfake, digital media governance, fake news, response mechanisms, social media.*

1. Đặt vấn đề

Những tiến bộ vượt bậc của trí tuệ nhân tạo (AI) trong những năm gần đây không chỉ tái định nghĩa các giới hạn của sáng tạo nội dung mà còn đặt ra nhiều vấn đề nan giải. Trong số đó, công nghệ deepfake một mặt là tiềm năng đổi mới, mặt khác là nguy cơ khuếch đại sự giả mạo, xuyên tạc và thao túng thông tin ở phạm vi rộng. Hệ quả là môi trường truyền thông toàn cầu phải đối mặt với vấn nạn thông tin giả mạo có quy mô và tốc độ lan truyền vượt xa các biện pháp kiểm soát truyền thống. Điều đáng lo ngại hơn cả là mạng xã hội, đặc biệt là ứng dụng Facebook của Meta, tính đến tháng 10/2024, Facebook tại Việt Nam ước tính có 86,1 triệu người dùng, chiếm 84% dân số [14] đã trở thành điểm nóng để deepfake bùng phát. Đặc biệt tại Việt Nam, vai trò của mạng xã hội như một nguồn tin thay thế càng củng cố khả năng ảnh hưởng sâu rộng của các sản phẩm deepfake đến nhận thức cộng đồng, làm xói mòn lòng tin vào các nguồn tin chính thống và thậm chí tác động đến các quyết định xã hội hoặc kích động các phản ứng tiêu cực [12].

Theo thống kê của các cơ quan chuyên môn, năm 2023 ghi nhận sự gia tăng đột biến số vụ việc liên quan đến deepfake và tin giả trên không gian mạng Việt Nam [3]. Mặc dù tác động và mức độ lan rộng của deepfake trên Facebook là vấn đề cảnh báo đỏ, song các nghiên cứu khoa học tại Việt Nam về chủ đề này vẫn chủ yếu đang ở mức nhận diện kỹ thuật hoặc khảo sát hiện tượng đơn lẻ, vắng bóng các phân tích hệ thống về mạng lưới lan truyền, động lực phát tán và tính hiệu quả của các giải pháp kiểm soát thực tiễn. Khoảng trống này dẫn đến sự thiếu hụt cơ sở khoa học cho công tác quản lý truyền thông số, xây dựng chính sách phòng ngừa, cũng như tổ chức các hoạt động giáo dục, truyền thông nâng cao nhận thức xã hội. Trong bối cảnh đó, việc thực hiện một nghiên cứu đánh giá về cơ chế ứng

Email: vuthiminhtam@gmail.com

phó với deepfake trên mạng xã hội ở Việt Nam là yêu cầu cấp thiết, vừa có ý nghĩa học thuật vừa mang giá trị thực tiễn sâu sắc. Về mặt thực tiễn, các phát hiện và khuyến nghị từ nghiên cứu sẽ cung cấp nền tảng vững chắc cho các nhà hoạch định chính sách, các tổ chức truyền thông, các nền tảng mạng xã hội cũng như cộng đồng xã hội trong việc xây dựng chiến lược phòng ngừa, nhận diện và ứng phó với deepfake một cách chủ động, hiệu quả.

2. Nội dung nghiên cứu

2.1. Tổng quan các nghiên cứu liên quan

Các mô hình lý thuyết, thực nghiệm cùng các giải pháp can thiệp đã được phát triển đa chiều, tạo nền tảng cho nhận thức và ứng phó với những thách thức do thông tin sai lệch gây ra. Nổi bật, Jin và cộng sự (2013) ứng dụng mô hình SIS trên Twitter để mô phỏng sự lây lan của tin giả. Nghiên cứu cho thấy quá trình lan truyền thông tin sai lệch có thể được mô hình hóa tương tự dịch bệnh khi một người dùng khi tiếp nhận tin giả có thể chuyển từ trạng thái dễ bị ảnh hưởng sang "bị nhiễm", sau đó hoặc dừng chia sẻ hoặc tiếp tục khuếch tán tới mạng lưới của mình [8]. Điểm mạnh của tiếp cận này là cho phép định lượng tốc độ, phạm vi và chu kỳ lan truyền, đồng thời dự báo khả năng bùng phát khi hội tụ đủ điều kiện về cấu trúc mạng lưới và cường độ tương tác. Các nghiên cứu của Borge-Holthoefer và cộng sự (2013) và Friggeri và cộng sự (2014) cho thấy cách thông tin lan truyền qua nhiều lớp người dùng trên Facebook giúp hiểu rõ hơn tầm quan trọng của cấu trúc mạng xã hội trong việc phát tán tin giả. Khi một nội dung nhận được sự quan tâm ban đầu của những nút trung tâm, hiệu ứng lan tỏa dây chuyền được kích hoạt và khuếch đại, dẫn tới hiện tượng thông tin lan rộng theo cấp số nhân trong thời gian ngắn [2], [6].

Một hướng nghiên cứu quan trọng khác tập trung vào cơ chế tác động thúc đẩy hoặc hạn chế sự lan truyền deepfake và tin giả. Allcott & Gentzkow (2017) nghiên cứu về ảnh hưởng của tin giả trong cuộc bầu cử Mỹ 2016 đã phân tích kỹ thuật thuật toán lan truyền của Facebook và cho thấy các thuật toán này được thiết kế để tối ưu hóa sự chú ý và tương tác, vô tình trở thành động lực thúc đẩy nội dung sai lệch tiếp cận nhanh và rộng [1]. Nghiên cứu này được coi là bằng chứng thực nghiệm điển hình cho nhận định về yếu tố kỹ thuật nền tảng mạng xã hội, kết hợp với hiệu ứng lan truyền đám đông đã tạo ra điều kiện thuận lợi cho deepfake và tin giả chiếm ưu thế. Vosoughi, Roy & Aral (2018 khẳng định các yếu tố như quan điểm chủ quan, động cơ chia sẻ dựa vào cảm xúc, sự tò mò hoặc thiếu kỹ năng kiểm chứng là những chất xúc tác then chốt khiến tin giả, deepfake phát tán mạnh hơn thông tin xác thực [15]. Bên cạnh đó, công trình của Del Vicario và cộng sự (2016) đã chứng minh sự hình thành các cộng đồng bong bóng thông tin khi người dùng chỉ tương tác với nhóm đồng thuận về quan điểm, càng làm gia tăng hiệu ứng cộng hưởng thông tin sai lệch và giảm hiệu quả phản biện xã hội [5]. Đây là vấn đề đặc biệt nghiêm trọng với deepfake, vốn có thể được làm bằng chứng trực quan cho những gì nhóm cộng đồng muốn tin, thay vì dựa vào xác thực khách quan.

2.2. Đánh giá hiệu quả các cơ chế ứng phó hiện nay đối với deepfake tại Việt Nam

2.2.1. Cơ chế cảnh báo

Cơ chế cảnh báo hiện là tuyến phòng vệ đầu tiên của các nền tảng số trước sự lan truyền của deepfake. Các cơ chế này được thiết kế nhằm phát hiện, gắn nhãn và chủ động cảnh báo người dùng về những nội dung nghi ngờ là giả mạo, qua đó giúp giảm thiểu nguy cơ chia sẻ và phát tán. Trên thực tế, Facebook đã đồng loạt triển khai nhiều hình thức cảnh báo hoặc liên kết đến bài viết kiểm chứng của các tổ chức fact-check độc lập. Đặc biệt, Twitter/X còn phát triển hệ thống "Community Notes" cho phép cộng đồng người dùng bổ sung, hiệu đính hoặc xác nhận thông tin của một bài đăng, từ đó tạo lớp cảnh báo đa chiều và gần với tâm lý người dùng hơn [4]. Tuy nhiên, khả năng phát hiện tự động các nội dung deepfake vẫn chịu nhiều hạn chế, nhất là khi deepfake ngày càng tinh vi, thường xuyên thay đổi định dạng hoặc giảm chất lượng nhằm qua mặt thuật toán. Trong khi đó, mạng lưới đối tác fact-check dù có kinh nghiệm chuyên môn lại không đáp ứng kịp về tốc độ lẫn quy mô, đặc biệt tại các thị trường ngoài Mỹ hoặc châu Âu. Việc Meta cắt giảm hợp tác fact-check ở Mỹ năm 2025 cũng cho thấy

sự thiếu bền vững khi phụ thuộc vào bên thứ ba và đặt ra nguy cơ giảm hiệu quả cảnh báo trên toàn cầu [13]. Các nghiên cứu gần đây khẳng định nhãn cảnh báo có thể làm giảm đáng kể tỷ lệ chia sẻ tin giả trong ngắn hạn, tuy nhiên, hiệu ứng này thường sẽ giảm tác dụng theo thời gian, thậm chí phản tác dụng với các nhóm người dùng đã có định kiến hoặc nằm trong các cộng đồng khép kín. Một số người dùng xem cảnh báo như biểu hiện kiểm duyệt hoặc tệ hơn coi đó là bằng chứng thông tin bị che giấu, từ đó tăng xu hướng chia sẻ nội dung sai lệch như một hình thức phản kháng [11].

Bên cạnh đó, thuật toán phân phối nội dung trên các nền tảng số hiện vẫn ưu tiên chỉ số tương tác như lượt thích, chia sẻ, bình luận hoặc thời lượng xem hơn tiêu chí xác thực. Điều này khiến nhiều nội dung deepfake hoặc tin giả được lan tỏa rộng rãi trước khi bị gắn nhãn cảnh báo. Đáng lưu ý, các tương tác mang tính tiêu cực như tranh luận, phẫn nộ, cảnh báo vẫn được hệ thống xem là tín hiệu tăng khả năng tiếp cận, tạo ra vòng lặp khuếch đại bất lợi cho các giải pháp kiểm soát [1]. Mặt khác, độ trễ trong việc phát hiện và cảnh báo các nội dung deepfake mới phát sinh. Khoảng thời gian từ khi một nội dung được đăng tải đến khi được gắn nhãn kiểm chứng có thể kéo dài từ vài giờ đến vài ngày, trong khi tốc độ lan truyền tự nhiên của mạng xã hội có thể đạt đến hàng trăm nghìn lượt xem hoặc chia sẻ chỉ trong thời gian rất ngắn. Các rào cản ngôn ngữ, văn hóa và thiếu hụt nguồn lực kiểm duyệt càng tăng nguy cơ deepfake xâm nhập mà không gặp trở ngại.

Nghịch lý tồn tại ở việc trong khi các nỗ lực gắn nhãn cảnh báo được thực thi, các thuật toán nền tảng lại ưu tiên phân phối những nội dung có tính tương tác cao, vốn là đặc tính tự nhiên của deepfake. Tại Việt Nam, hiệu quả của nhãn cảnh báo đối mặt với rào cản từ tâm lý hoài nghi của người sử dụng, dẫn đến xu hướng chia sẻ nội dung như một hình thức phản kháng thông tin chính thống. Việc thiếu hụt sự thấu hiểu về đặc thù hành vi và bối cảnh văn hóa truyền thông khiến các giải pháp kỹ thuật thuần túy chưa đạt được mục tiêu điều chỉnh nhận thức cộng đồng một cách bền vững.

2.2.2. Cơ chế nhận diện và kiểm chứng

Cơ chế nhận diện và kiểm chứng deepfake đang phát triển mạnh mẽ với sự kết hợp giữa công nghệ học sâu và các giải pháp xã hội. Trong môi trường thực tiễn, đặc biệt tại Việt Nam, việc triển khai các giải pháp này lại đặt ra nhiều thách thức về hiệu quả, tính ứng dụng và sự phối hợp liên ngành. Các mô hình học sâu như XceptionNet, EfficientNet, MesoNet hay các hệ thống phân tích chuyển động, phát hiện bất thường vi mô trên khuôn mặt, phân tích tín hiệu âm thanh đều cho kết quả nổi bật về độ chính xác trong các thử nghiệm chuẩn quốc tế [4], [13]. Một số tổ chức quốc tế cũng bắt đầu phát triển hệ thống phát hiện đa phương tiện áp dụng ở quy mô nền tảng, kết hợp phân tích hình ảnh, video, âm thanh và văn bản, giúp tăng năng lực phòng vệ trước deepfake thế hệ mới. Tuy nhiên, khi áp dụng trong môi trường thực tế của mạng xã hội, đặc biệt là Facebook, nền tảng phổ biến tại Việt Nam hiệu quả nhận diện bị suy giảm rõ rệt. Các video deepfake được chia sẻ nhiều vòng, liên tục chỉnh sửa định dạng hoặc cắt ghép lồng ghép vào meme làm giảm đáng kể năng lực phát hiện của các thuật toán hiện hành.

Ở Việt Nam, hệ thống nhận diện tự động chủ yếu dựa trên giải pháp từ các module mã nguồn mở, chưa xuất hiện các hệ thống kiểm chứng do doanh nghiệp hoặc viện nghiên cứu trong nước phát triển và làm chủ hoàn toàn. Các chuyên gia an ninh mạng tại Việt Nam cũng nhiều lần cảnh báo về lỗ hổng trong năng lực nhận diện khi deepfake sử dụng giọng nói tiếng Việt hoặc các đặc điểm nhận diện không phổ biến trên các bộ dữ liệu quốc tế. Hiện tại, Việt Nam chủ yếu dựa vào các tổ chức kiểm chứng độc lập như VFC (Vietnam Fact-Checking), các phòng kiểm chứng của một số cơ quan báo chí lớn (VTV, VnExpress), các sáng kiến hợp tác với tổ chức quốc tế như AFP Fact Check, Google Fact Check. Mặc dù vậy, số lượng nhân sự kiểm chứng còn mỏng, các quy trình xác minh và cảnh báo chủ yếu theo phương thức thủ công, chưa có sự phối hợp hoặc chia sẻ dữ liệu đồng bộ giữa các bên. Hơn nữa, thiếu tiêu chuẩn kiểm chứng liên nền tảng là một điểm yếu. Một video deepfake có thể bị gỡ bỏ trên Facebook sau khi có xác minh, nhưng vẫn lan truyền tự do trên các nền tảng khác hoặc qua các nhóm chat kín mà không bị chặn kịp thời. Điều này làm giảm hiệu quả kiểm soát tổng thể và tạo điều kiện cho deepfake dưới dạng biến thể hoặc lây lan chéo giữa các hệ sinh thái.

2.2.3. Cơ chế xử lý và can thiệp xã hội

Không chỉ dừng lại ở phát hiện và cảnh báo, việc xử lý hiệu quả cần kết hợp đồng bộ giữa các biện pháp kỹ thuật, chính sách pháp lý và giải pháp xã hội nhằm ngăn chặn, giảm thiểu tác động tiêu cực, cũng như phục hồi lòng tin vào hệ sinh thái thông tin số. Nhiều quốc gia và tổ chức đã chủ động hoàn thiện hành lang pháp lý, khung xử lý và hợp tác đa ngành nhằm đối phó với nguy cơ lan truyền của deepfake. Liên minh châu Âu triển khai "Code of Practice on Disinformation" và tiến tới áp dụng Digital Services Act (DSA), yêu cầu nền tảng lớn như Facebook, YouTube, TikTok… phải nhanh chóng gỡ bỏ hoặc hạn chế tiếp cận đối với các nội dung xác định là giả mạo, đồng thời công bố minh bạch các biện pháp can thiệp, mức độ hợp tác với tổ chức kiểm chứng và cơ quan quản lý nhà nước [4]. Tại Mỹ và nhiều quốc gia phát triển, việc sử dụng deepfake với mục đích gian lận, xúc phạm cá nhân hoặc phá hoại an ninh quốc gia đã bị xếp vào nhóm hành vi vi phạm nghiêm trọng, có thể bị truy tố hình sự, xử phạt hành chính nặng hoặc buộc bồi thường dân sự. Tuy nhiên, hiệu quả thực tế của các cơ chế xử lý này vẫn chịu tác động mạnh bởi đặc thù pháp lý và văn hóa từng quốc gia.

Tại Việt Nam, Chính phủ đã ban hành Luật An ninh mạng năm 2018, đặt nền tảng cho quản lý hoạt động trên không gian mạng, buộc các nền tảng cung cấp dịch vụ mạng xã hội tại Việt Nam phải kiểm soát, ngăn chặn và phối hợp xử lý các nội dung vi phạm. Nghị định 15/2020/NĐ-CP quy định chi tiết xử phạt hành chính trong lĩnh vực bưu chính, viễn thông, công nghệ thông tin và giao dịch điện tử. Nghị định 53/2022/NĐ-CP hướng dẫn thi hành một số điều của Luật An ninh mạng, trong đó tăng cường yêu cầu lưu trữ, kiểm soát dữ liệu, xác minh chủ tài khoản và phối hợp gỡ bỏ nội dung độc hại theo yêu cầu cơ quan chức năng. Ngoài ra, Thông tư 38/2016/TT-BTTTT cũng quy định chi tiết về việc cung cấp thông tin công cộng qua biên giới. Các văn bản này nhấn mạnh trách nhiệm pháp lý của cá nhân, tổ chức và nhà cung cấp nền tảng trong kiểm soát, phát hiện và xử lý thông tin sai lệch, deepfake trên không gian mạng. Tuy nhiên, đánh giá khách quan cho thấy các vụ deepfake ở Việt Nam hiện chủ yếu dừng lại ở mức độ xử lý hành chính, gỡ bỏ nội dung, yêu cầu đính chính hoặc nhắc nhở, hiếm khi bị xử phạt nặng hoặc truy cứu trách nhiệm hình sự, đặc biệt khi thủ phạm ở ngoài lãnh thổ Việt Nam. Hạn chế lớn là quy trình xác minh, chuyển giao xử lý giữa các bộ ngành còn thiếu đồng bộ, chưa xây dựng được quy trình phản ứng nhanh khi phát hiện vụ việc nghiêm trọng.

Dù nhiều chiến dịch truyền thông cảnh báo đã được các cơ quan báo chí, tổ chức kiểm chứng, các đài truyền hình lớn phối hợp triển khai, song chưa có chương trình giáo dục chính khóa về kỹ năng số, kỹ năng kiểm chứng hoặc nhận diện thông tin giả mạo trong hệ thống giáo dục phổ thông. Đa số người dùng mạng xã hội Việt Nam vẫn chưa hình thành thói quen kiểm chứng nguồn tin, dễ bị ảnh hưởng bởi tâm lý đám đông, hiệu ứng lan truyền nhanh, hoặc thao túng từ các nhóm kín. Khả năng truy xuất, lưu vết và chia sẻ dữ liệu nhận diện, xử lý giữa các nền tảng và cơ quan quản lý còn phân mảnh, chưa thống nhất chuẩn chung về xác minh, phản hồi và xử lý vụ việc. Điều này tạo điều kiện cho deepfake di chuyển, biến tướng và lẩn tránh qua nhiều nền tảng, gây khó khăn lớn cho việc kiểm soát hiệu quả ở cấp hệ sinh thái.

2.3. Đề xuất khuyến nghị liên quan

2.3.1. Phát triển công nghệ nhận diện deepfake mang tính bản địa hóa, lấy nhu cầu thực tiễn Việt Nam làm trung tâm

Vấn đề quan trọng hiện nay là phải xây dựng nền tảng công nghệ nhận diện deepfake dựa trên cơ sở dữ liệu bản địa bao gồm ngôn ngữ, khuôn mặt, giọng nói, thói quen giao tiếp, đặc điểm văn hóa của người Việt. Để hiện thực hóa, cần phát động các chương trình nghiên cứu và phát triển giữa các trường đại học, viện công nghệ, doanh nghiệp công nghệ thông tin lớn, với nguồn tài trợ và bảo trợ của Nhà nước. Giai đoạn đầu có thể là việc thiết lập các nhóm chuyên gia xây dựng kho dữ liệu ảnh, video, âm thanh, meme phổ biến trong môi trường mạng Việt Nam; đào tạo các mô hình học sâu nhận diện deepfake trên nền dữ liệu này, liên tục cập nhật các mẫu giả mạo mới phát sinh từ thực tiễn. Các giải pháp nhận diện cần được thử nghiệm rộng rãi trên các nền tảng mạng xã hội, lấy phản hồi từ cộng

đồng để hiệu chỉnh và bổ sung. Không chỉ dừng lại ở phát triển thuật toán, các nền tảng công nghệ lớn hoạt động tại Việt Nam cũng phải có trách nhiệm phối hợp, tích hợp giải pháp phát hiện và cảnh báo tự động này vào hệ thống kiểm duyệt nội dung, tạo nên bức tường phòng vệ chủ động ngay từ khâu tiếp nhận thông tin.

2.3.2. Hình thành một hệ sinh thái kiểm chứng đa tầng, liên kết chặt chẽ giữa các bộ ngành, tổ chức kiểm chứng, doanh nghiệp công nghệ và các nền tảng số

Việc kiểm chứng deepfake và tin giả hiện nay ở Việt Nam cần có sự kết nối và các tiêu chuẩn chia sẻ dữ liệu xuyên nền tảng. Đề xuất đặt trọng tâm vào việc thành lập một trung tâm dữ liệu kiểm chứng quốc gia, nơi tích hợp cảnh báo, dữ liệu và kết quả kiểm chứng từ Bộ Khoa học và Công nghệ, Bộ Công an, các doanh nghiệp viễn thông, tổ chức fact-check độc lập, các nền tảng mạng xã hội và cơ quan báo chí. Trung tâm này vận hành trên cơ chế chia sẻ cảnh báo theo thời gian thực, sử dụng bộ tiêu chuẩn, quy trình xác minh và nhãn cảnh báo chung, đảm bảo khi một nội dung bị xác định là deepfake ở một nền tảng, dữ liệu sẽ tự động đồng bộ hóa, cảnh báo và gỡ bỏ trên các nền tảng khác. Để đạt được điều này thì cần một đội ngũ kiểm chứng viên có chuyên môn sâu, phát triển công nghệ xác minh mới và điều phối phản ứng nhanh khi xuất hiện sự cố deepfake nguy hiểm trên diện rộng. Hệ sinh thái này vừa bảo đảm kiểm soát hiệu quả, vừa tăng tính minh bạch, tạo niềm tin cho xã hội và cộng đồng sử dụng mạng.

2.3.3. Tăng cường sức mạnh pháp lý và chế tài răn đe thực chất đối với các hành vi phát tán deepfake và tin giả

Để khắc phục, cần chủ động hoàn thiện các văn bản hướng dẫn xác định, định danh hành vi deepfake nguy hiểm, xây dựng quy trình điều tra truy vết số có sự phối hợp chặt chẽ giữa Bộ Công an, Bộ Khoa học và Công nghệ, tòa án và các nền tảng mạng xã hội. Mức xử phạt hành chính đối với hành vi phát tán, sản xuất, tiếp tay lan truyền deepfake nguy hiểm cần được nâng cao, bổ sung hình phạt bổ sung. Quy trình xử lý phải minh bạch, công bố rộng rãi, lấy trường hợp điển hình để cảnh báo, răn đe. Đặc biệt, cần bổ sung quy định trách nhiệm của các nền tảng xuyên biên giới trong việc phối hợp cung cấp dữ liệu truy xuất nguồn gốc, gỡ bỏ nội dung deepfake theo yêu cầu từ cơ quan chức năng Việt Nam trong thời hạn tối đa, tránh việc deepfake lẩn tránh bằng cách chuyển nền tảng hoặc ẩn danh trên không gian số.

2.3.4. Đẩy mạnh giáo dục truyền thông số, lấy cộng đồng làm trung tâm để chủ động phòng vệ trước deepfake

Truyền thông số và kỹ năng nhận diện thông tin giả mạo phải trở thành một phần không thể thiếu trong chương trình giáo dục từ phổ thông đến đại học, thậm chí đào tạo thường xuyên cho cán bộ, công nhân viên chức và các nhóm dễ bị tổn thương (người cao tuổi, người lao động phổ thông). Lồng ghép kiến thức nhận diện deepfake, kỹ năng kiểm chứng thông tin, kỹ năng bảo vệ bản thân trước tin giả vào giáo trình các môn học xã hội, giáo dục công dân, ngữ văn; xây dựng các module học trực tuyến miễn phí cho cộng đồng. Có thể tổ chức cuộc thi, trải nghiệm thực tế, diễn tập phòng vệ tin giả ngay tại các trường học, doanh nghiệp, đoàn thể. Song song đó, các nền tảng mạng xã hội ở Việt Nam cần phối hợp với các cơ quan báo chí, các KOLs và các nhóm cộng đồng mạng để lan tỏa các câu chuyện thực tế, bài học cảnh báo và mô hình phản biện sáng tạo, giúp cộng đồng chủ động trước cạm bẫy thông tin giả, thay vì chỉ trông chờ vào công nghệ hay pháp luật.

3. Kết luận

Việc đánh giá các cơ chế hiện tại cho thấy một khoảng cách đáng kể giữa tốc độ phát triển của công nghệ giả mạo và năng lực kiểm soát của các cơ quan quản lý cũng như các nền tảng mạng xã hội tại Việt Nam. Các giải pháp kỹ thuật như gắn nhãn cảnh báo hay thuật toán nhận diện mặc dù cần thiết nhưng thường xuyên rơi vào trạng thái bị động trước sự biến đổi nhanh chóng của các kịch bản giả mạo và sự thiếu hụt dữ liệu huấn luyện mang tính bản địa. Để ứng phó hiệu quả, việc chuyển dịch từ mô hình xử lý tình huống sang chiến lược quản trị đa bên chủ động là yêu cầu cấp thiết. Thành công của quá trình này phụ thuộc vào sự phối hợp nhịp nhàng giữa bốn yếu tố: công nghệ nhận diện đặc

thù, hệ thống pháp luật minh bạch với chế tài đủ mạnh, sự gắn kết giữa các chủ thể trong hệ sinh thái thông tin và năng lực phản biện của người dùng. Trong đó, việc xây dựng hệ miễn dịch số thông qua giáo dục kỹ năng kiểm chứng thông tin chính là giải pháp gốc rễ để hạn chế sự lây lan của tin giả và phục hồi lòng tin của cộng đồng vào môi trường truyền thông số. Những khuyến nghị trong nghiên cứu này được đề xuất nhằm xây dựng một không gian mạng an toàn và bền vững hơn tại Việt Nam trong tương lai.

Tài liệu tham khảo

[1] Allcott, H., & Gentzkow, M. (2017). *Social media and fake news in the 2016 election*. Journal of Economic Perspectives, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211.

[2] Borge-Holthoefer, J., Baños, R. A., González-Bailón, S., & Moreno, Y. (2013). *Cascading behavior in complex socio-technical networks.* Journal of Complex Networks, 1(1), 3–24. https://doi.org/10.1093/comnet/cnt006.

[3] Chesney, R., & Citron, D. K. (2019). *Deep fakes: A looming challenge for privacy, democracy, and national security.* California Law Review, 107, 1753–1819. https://doi.org/10.2139/ssrn.3213954.

[4] Chuai, G., et al. (2024). *Community Notes on Social Media: Effectiveness and Challenges.* arXiv preprint. https://arxiv.org/abs/2409.08781.

[5] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). *Echo chambers: Emotional contagion and group polarization on Facebook*. Scientific Reports, 6, 37825. https://doi.org/10.1038/srep37825.

[6] Friggeri, A., Adamic, L. A., Eckles, D., & Cheng, J. (2014). *Rumor Cascades. Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM 2014).* https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8122.

[7] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial nets.* Advances in Neural Information Processing Systems, 27, 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[8] Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). *Epidemiological modeling of news and rumors on Twitter.* Proceedings of the 7th Workshop on Social Network Mining and Analysis, 1–9. https://doi.org/10.1145/2501025.2501027.

[9] Lao động (2023). *Deepfake và những chiêu lừa đảo tinh vi khiến nhiều người sập bẫy năm 2023.* Nguồn: https://laodong.vn/cong-nghe/deepfake-va-nhung-chieu-lua-dao-tinh-vi-khien-nhieu-nguoi-sap-bay-nam-2023-1283790.ldo.

[10] Pastor-Satorras, R., & Vespignani, A. (2001). *Epidemic spreading in scale-free networks.* Physical Review Letters, 86(14), 3200–3203. https://doi.org/10.1103/PhysRevLett.86.3200

[11] Pennycook, G., & Rand, D. G. (2020). *Fighting misinformation on social media using crowdsourced judgments of news source quality.* PNAS, 117(6), 2775–2783. https://doi.org/10.1073/pnas.1912444117.

[12] Sở Thông tin & Truyền thông Nghệ An (2025). *Cách chống lại hệ lụy sâu rộng từ deepfake.* Nguồn: https://naict.tttt.nghean.gov.vn/pckns/cach-chong-lai-he-luy-sau-rong-tu-deepfake-1618.html.

[13] The Guardian (2025). *Dispiriting: A factchecker reacts to Meta Facebook move to scrap role*. https://www.theguardian.com/technology/2025/jan/08/dispiriting-a-factchecker-reacts-to-meta-facebook-move-to-scrap-role.

[14] VnEconomy (2024). *Facebook users in Vietnam estimated at more than 86 million.* Nguồn: https://en.vneconomy.vn/facebook-users-in-vietnam-estimated-at-more-than-86-million.htm

[15] Vosoughi, S., Roy, D., & Aral, S. (2018). *The spread of true and false news online.* Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559.

[16] Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

[17] Westerlund, M. (2019). *The emergence of deepfake technology: A review.* Technology Innovation Management Review, 9(11), 40–53. https://doi.org/10.22215/timreview/1282.