Tạp chí Lý Luận Chính Trị Và Truyên Thông

Tạp chí điện tử tiếng Việt - Tập 32, Số 7 (2026)

Bảo vệ nền tảng tư tưởng của Đảng trước thách thức trí tuệ nhân tạo (Kỳ 2: Cơ chế tác động tâm lý - xã hội và kinh nghiệm quốc tế)

Ngày nhận bài: 27/04/2026 | **Ngày phản biện:** 08/05/2026 | **Ngày xuất bản:** 05/07/2026

Tác giả: Nguyễn Thị Ngọc Hoa (Học viện viện Báo chí và Tuyên truyền)

DOI: 10.66057/llcttt-29751

Tóm tắt: Kỳ 2 của loạt bài nghiên cứu về bảo vệ nền tảng tư tưởng của Đảng trước tấn công bằng trí tuệ nhân tạo (AI) tập trung phân tích hai vấn đề cốt lõi, mang tính chìa khóa để xây dựng chiến lược phòng thủ hiệu quả. Thứ nhất là cơ chế tác động tâm lý - xã hội của các thủ đoạn tấn công tư tưởng sử dụng AI, trọng tâm là ba cơ chế: khai thác thiên kiến nhận thức, tận dụng hiệu ứng buồng vọng và chiến lược nhắm mục tiêu vi mô dựa trên phân tích tâm lý bằng dữ liệu lớn. Thứ hai là kinh nghiệm quốc tế trong đối phó với thách thức tư tưởng số, bao gồm mô hình tổng thể của Liên minh châu Âu, kinh nghiệm pháp lý AI của Trung Quốc, mô hình cân bằng của Singapore và kinh nghiệm thực chiến của Hoa Kỳ. Trên cơ sở đó, bài viết chắt lọc những giá trị tham khảo phù hợp với bối cảnh Việt Nam, chuẩn bị cho hệ thống giải pháp sẽ được đề xuất ở Kỳ 3.

1. Đặt vấn đề

Kỳ 1 của bài viết đã phân tích năm nhóm thủ đoạn tấn công tư tưởng sử dụng AI trên không gian mạng Việt Nam. Kỳ 2 này tiếp cận câu hỏi sâu hơn và có tính quyết định hơn về mặt chiến lược: tại sao những thủ đoạn đó lại hiệu quả đến vậy? Điều gì trong cơ chế tâm lý con người khiến chúng ta dễ bị tổn thương trước tấn công tư tưởng bằng AI? Và tại sao AI lại có khả năng khai thác những điểm yếu đó hiệu quả hơn bất kỳ hình thức tấn công tư tưởng nào trước đây? Trả lời chính xác những câu hỏi này là điều kiện tiên quyết để xây dựng chiến lược phòng thủ đúng đắn, không thể "chữa bệnh" đúng cách nếu không "chẩn đoán đúng bệnh".

C.Mác khẳng định trong tác phẩm "Luận cương về Phoiơbắc" (1845): "Trong tính hiện thực của nó, bản chất con người là tổng hòa những quan hệ xã hội" (Mác & Ăngghen, 1995, t.3, tr.11). Luận điểm sâu sắc đó, trong thời đại AI, có một hệ quả quan trọng: khi các "quan hệ xã hội" của hàng chục triệu người ngày càng được trung gian hóa qua các thuật toán số, thì kẻ kiểm soát thuật toán cũng kiểm soát một phần quan trọng môi trường xã hội và qua đó tác động đến ý thức xã hội. Đây là lý do tại sao cuộc đấu tranh kiểm soát không gian số không phải là cuộc đấu tranh kỹ thuật thuần túy, mà là cuộc đấu tranh chính trị - tư tưởng có tầm quan trọng chiến lược.

Trong tác phẩm "Sửa đổi lối làm việc" (1947), Chủ tịch Hồ Chí Minh chỉ rõ: "Biết người cố nhiên là khó, tự biết mình cũng không dễ" (Hồ Chí Minh, 2011, t.5, tr.277). Chỉ dẫn phương pháp luận đó đặc biệt đúng trong cuộc đấu tranh tư tưởng hôm nay: muốn bảo vệ tư tưởng của cộng đồng một cách hiệu quả, trước hết phải hiểu chính cộng đồng đó, hiểu vì sao và bằng cách nào con người dễ bị tổn thương trước tấn công tư tưởng bằng AI. Trong thời đại AI, sự "hiểu biết" cần thiết đó nhất thiết phải bao gồm việc hiểu cơ chế tâm lý mà AI đang khai thác, đây là tiền đề của mọi chiến lược phòng thủ tư tưởng hiệu quả. Không thể bảo vệ tư tưởng của cộng đồng nếu không hiểu tại sao và như thế nào cộng đồng đó dễ bị tổn thương.

2. Phương pháp nghiên cứu

Về phương pháp nghiên cứu, bài viết dựa trên phương pháp luận của chủ nghĩa duy vật biện chứng và chủ nghĩa duy vật lịch sử, đặt hiện tượng tấn công tư tưởng bằng AI trong mối quan hệ biện chứng giữa tồn tại xã hội và ý thức xã hội. Trên nền tảng đó, Kỳ 2 sử dụng kết hợp các phương pháp nghiên cứu định tính: phân tích - tổng hợp tài liệu đối với các công trình tâm lý học nhận thức cùng các báo cáo, văn bản pháp lý trong và ngoài nước; phương pháp liên ngành kết hợp chính trị học - tâm lý học xã hội - truyền thông học - khoa học máy tính để lý giải cơ chế tác động; và đặc biệt là phương pháp so sánh để khảo sát kinh nghiệm của một số quốc gia, tổ chức, từ đó chắt lọc những giá trị có thể tham khảo cho Việt Nam. Mục tiêu xuyên suốt của Kỳ 2 là cung cấp luận cứ khoa học cho việc kiến tạo "lá chắn tư tưởng số" bảo vệ vững chắc nền tảng tư tưởng của Đảng trên không gian mạng trong kỷ nguyên phát triển mới: muốn dựng được lá chắn ấy, trước hết phải hiểu thấu cơ chế tấn công (nội dung Phần II) và tham khảo có chọn lọc kinh nghiệm ứng phó của thế giới (nội dung Phần III).

3. Kết quả và thảo luận

3.1. Kết quả nghiên cứu

3.1.1. Cơ chế tác động tâm lý - xã hội của tấn công tư tưởng bằng AI

3.1.1.1. Khai thác thiên kiến nhận thức, điểm yếu cấu trúc của tư duy con người

Khoa học tâm lý học nhận thức hiện đại, đặc biệt từ những nghiên cứu tiên phong của Daniel Kahneman và Amos Tversky từ những năm 1970 đến nay, đã chứng minh có hệ thống: con người không phải là "máy tính lý trí" như triết học cổ điển và kinh tế học tân cổ điển từng giả định. Thay vào đó, con người sở hữu hàng chục loại thiên kiến nhận thức, nói một cách dễ hiểu, đó là những "lối mòn", những "đường tắt" trong tư duy giúp não bộ phán đoán nhanh, nhưng lại dẫn đến sai lầm một cách có hệ thống và có thể dự đoán được trong những điều kiện nhất định. Chính vì có thể dự đoán được, các thiên kiến này trở thành "điểm yếu cấu trúc" để AI khai thác.

Nhà tâm lý học và kinh tế học đoạt giải Nobel Daniel Kahneman (2011, tr.59-70) mô tả hai hệ thống tư duy song song của con người: Hệ thống 1 (nhanh, bản năng, cảm xúc, không cần nỗ lực ý thức, dễ bị thiên kiến) và Hệ thống 2 (chậm, lý trí, phân tích, đòi hỏi nỗ lực ý thức, khó bị thiên kiến hơn). Phần lớn các quyết định trong cuộc sống hằng ngày, đặc biệt là quyết định có chia sẻ hay tin vào một thông tin trên mạng xã hội hay không, được thực hiện bởi Hệ thống 1 mà không qua Hệ thống 2. AI tấn công tư tưởng được thiết kế chính xác để nhắm vào Hệ thống 1: tạo ra nội dung kích hoạt phản ứng cảm xúc tức thì (phẫn nộ, sợ hãi, nghi ngờ, bất bình) trước khi Hệ thống 2 kịp phân tích. Một khi Hệ thống 1 đã quyết định "chia sẻ", Hệ thống 2 khó mà ngăn cản.

Nhà tâm lý học Jonathan Haidt (2012, tr.55-60) chỉ ra thêm: phần lớn các phán đoán chính trị và đạo đức của con người được hình thành bởi cảm xúc và trực giác trước, lý trí chỉ đến sau để hợp lý hóa. Điều đó có nghĩa: nội dung tấn công tư tưởng hiệu quả nhất không phải là nội dung "thuyết phục bằng lý luận" mà là nội dung "kích hoạt cảm xúc" và AI đặc biệt giỏi trong việc tạo ra nội dung kích hoạt cảm xúc đúng loại, với đúng đối tượng, vào đúng thời điểm.

Ba thiên kiến nhận thức bị khai thác nhiều nhất trong các chiến dịch tấn công tư tưởng bằng AI là: (1) Thiên kiến xác nhận, nghĩa là con người có xu hướng chỉ tìm kiếm, tin và ghi nhớ những thông tin xác nhận điều mình vốn đã tin, đồng thời bỏ qua hoặc bác bỏ những thông tin trái chiều; thuật toán AI khai thác điểm này bằng cách liên tục "rót" vào người dùng đúng loại nội dung hợp với định kiến của họ, từng bước kéo họ vào vòng xoáy ngày càng cực đoan; (2) Hiệu ứng sự thật ảo giác, nghĩa là một thông tin nếu được nghe đi nghe lại nhiều lần sẽ dần được não bộ cảm nhận là "đáng tin", dù nó chưa hề được kiểm chứng; mạng lưới tài khoản tự động (bot) có thể lặp lại một tin giả hàng nghìn lần mỗi ngày, tạo ra hiệu ứng này ở quy mô chưa từng có; (3) Hiệu ứng bằng chứng xã hội, nghĩa là con người có xu hướng tin và làm theo số đông, nhất là khi bản thân đang phân vân. Nhà tâm lý học xã hội Solomon Asch (1955, tr.31-35) đã chứng minh trong loạt thí nghiệm kinh điển rằng người ta sẵn sàng đồng ý với một đáp án rõ ràng là sai chỉ vì thấy đa số người khác chọn đáp án đó. Mạng lưới bot do AI điều khiển khai thác hiệu ứng này khi hàng nghìn tài khoản đồng loạt "ủng hộ" một luận điệu thù địch, tạo ảo giác rằng "ai cũng nghĩ vậy".

3.1.1.2. Hiệu ứng buồng vọng, vũ khí tâm lý có hệ thống của thuật toán phân cực

Trong tác phẩm "Influence" (Ảnh hưởng và sức mạnh thuyết phục), Robert Cialdini (2006, tr.87-100) chỉ ra sáu nguyên tắc ảnh hưởng xã hội mà con người đặc biệt nhạy cảm, nổi bật là nguyên tắc "bằng chứng xã hội" và "yêu thích". Các thuật toán AI của mạng xã hội khai thác những nguyên tắc đó để tạo ra "buồng vọng" (echo chamber): không gian thông tin khép kín mà người dùng chỉ tiếp nhận những quan điểm củng cố thêm niềm tin sẵn có của họ, hầu như không gặp những quan điểm đối lập đủ thuyết phục.

Trong cuốn "The Filter Bubble" (Bong bóng lọc), Eli Pariser (2011, tr.74-80) đã đặt nền móng lý thuyết cho hiểu biết về nguy cơ này khi Internet mới bắt đầu bùng nổ - cảnh báo rằng các thuật toán gợi ý đang "ẩn" thế giới thực khỏi người dùng và tạo ra những "bong bóng lọc" ngày càng cô lập. Nhưng với sự ra đời của AI tạo sinh thế hệ mới và các thuật toán gợi ý tiên tiến hơn, "bong bóng lọc" đó đã trở nên nguy hiểm gấp bội: thuật toán không chỉ lọc thông tin mà còn chủ động tạo ra nội dung mới phù hợp với "bong bóng" của từng người dùng, tạo ra một vòng phản hồi ngày càng cực đoan hơn - người dùng tiếp nhận nội dung phù hợp thiên kiến, thiên kiến được củng

cố, thuật toán lại gợi ý nội dung cực đoan hơn, và thiên kiến càng sâu hơn.

Trong bối cảnh Việt Nam, hiệu ứng buồng vọng kết hợp với tâm lý "ngại đối đầu công khai" và xu hướng "tin vào nguồn quen" tạo ra nguy cơ đặc biệt: người dùng thường không phản bác công khai thông tin giả trên mạng dù trong lòng nghi ngờ, tạo ra cảm giác chấp nhận ngầm đối với thông tin sai lệch. Kết quả là các luận điệu thù địch về kinh tế, tham nhũng, dân chủ và nhân quyền có thể lan rộng mà không bị phản bác công khai tạo ra một "dư luận ngầm" tiêu cực về Đảng và Nhà nước ngay cả trong những người nhìn chung vẫn có niềm tin vào hệ thống. Đây là sự bào mòn chậm nhưng sâu, nguy hiểm hơn nhiều so với những cuộc tấn công ồn ào, dễ nhận biết và dễ phản bác.

3.1.1.3. Nhắm mục tiêu vi mô, cá nhân hóa tấn công dựa trên tâm lý học dữ liệu lớn

Nếu buồng vọng là vũ khí tấn công "diện rộng" nhắm vào một nhóm đối tượng lớn với thông điệp chung thì nhắm mục tiêu vi mô là vũ khí "điểm", nhắm vào từng cá nhân cụ thể với thông điệp được tùy chỉnh theo tâm lý riêng của từng người. Đây là bước tiến chưa có tiền lệ trong lịch sử chiến tranh tâm lý và tuyên truyền: lần đầu tiên, kẻ tấn công tư tưởng có thể "đo ni đóng giày" thông điệp cho hàng triệu cá nhân cùng một lúc, tự động và với chi phí gần bằng không.

Theo Pew Research Center (2024), người dùng mạng xã hội ngày càng khó phân biệt nội dung do AI tạo ra với nội dung do con người viết, nhất là khi chất lượng deepfake ngày càng cao. Reuters Institute for the Study of Journalism (2024, tr.12-16) ghi nhận một xu hướng đáng lo ngại: niềm tin của công chúng vào các thể chế chính thống đang suy giảm ở nhiều quốc gia mà một phần nguyên nhân được cho là tác động kéo dài của nhắm mục tiêu vi mô tiêu cực.

Quy trình nhắm mục tiêu vi mô về tư tưởng sử dụng AI bao gồm ba giai đoạn liên tiếp. Giai đoạn 1 thu thập và phân tích dữ liệu: AI phân tích hàng nghìn điểm dữ liệu của từng người dùng (lịch sử tìm kiếm, nội dung đã thích, thời gian sử dụng, mạng lưới quan hệ, thậm chí tốc độ cuộn màn hình và thời điểm dừng lại để đọc) nhằm dựng "hồ sơ tâm lý" chi tiết của họ. Giai đoạn 2 phân loại theo điểm yếu tâm lý: AI sắp xếp người dùng theo những điểm yếu có thể khai thác (lo lắng về kinh tế, bất bình trước bất công xã hội, hoài nghi thể chế, v.v.). Giai đoạn 3 gửi thông điệp cá nhân hóa: AI tạo ra thông

điệp được "may đo" theo hồ sơ tâm lý của từng người và gửi đến đúng thời điểm họ dễ tiếp nhận nhất. Nguy hiểm nhất là kỹ thuật này tạo ra ảo giác "tự mình nhận ra sự thật": người nhận thông điệp ngỡ rằng mình đang tự suy nghĩ và rút ra kết luận độc lập, mà không hề biết toàn bộ hành trình nhận thức đó đã được thiết kế từ trước.

Hiểu rõ ba cơ chế tâm lý - xã hội nêu trên chính là điều kiện tiên quyết để thiết kế "lá chắn tư tưởng số" đúng hướng: một lá chắn không chỉ ngăn chặn nội dung độc hại về mặt kỹ thuật, mà quan trọng hơn, phải vô hiệu hóa được chính cơ chế khai thác tâm lý mà AI đang dùng để xâm nhập nhận thức của quần chúng. Cũng chính vì vậy, việc tham khảo cách thức một số quốc gia, tổ chức trên thế giới ứng phó với thách thức này trở nên cần thiết.

3.2. Kinh nghiệm quốc tế trong đối phó với thách thức tư tưởng số

3.2.1. Liên minh châu Âu, chiến lược quản trị AI và thông tin tổng thể, toàn diện

Liên minh châu Âu (EU) là một trong những chủ thể dẫn đầu thế giới về xây dựng khung pháp lý và thể chế toàn diện để đối phó với thách thức thông tin giả và AI trong lĩnh vực dân chủ - chính trị. Cách tiếp cận của EU có ba đặc điểm nổi bật đáng học hỏi: tính hệ thống (kết hợp pháp lý, kỹ thuật và truyền thông trong một chiến lược nhất quán), tính liên tục (được điều chỉnh và nâng cấp theo sự thay đổi của công nghệ) và tính minh bạch (công khai hóa phương pháp và kết quả để tạo niềm tin xã hội).
Mạng lưới "EUvsDisinfo" của Cơ quan Đối ngoại châu Âu (EEAS), với đội ngũ chuyên gia phân tích đa ngôn ngữ hoạt động liên tục, mỗi năm ghi nhận và xử lý hàng nghìn chiến dịch thông tin sai lệch, trong đó ngày càng nhiều chiến dịch có sử dụng công nghệ AI (EEAS, 2024). Điều làm cho mô hình này hiệu quả là sự kết hợp tối ưu: AI đảm nhận việc sàng lọc và phân tích ban đầu (xử lý khối lượng nội dung khổng lồ mỗi ngày), trong khi chuyên gia con người đưa ra phán quyết cuối cùng (đảm bảo độ tin cậy và tính pháp lý). Kế hoạch hành động chống thông tin giả (European Commission, 2018, tr.2-8) và Đạo luật Trí tuệ nhân tạo (Quy định (EU) 2024/1689) quy định nhiều nghĩa vụ bắt buộc: nội dung do AI tạo ra phải được gắn nhãn rõ ràng; các mô hình AI có rủi ro cao phải trải qua đánh giá độc lập trước khi triển khai; người dùng có quyền được biết khi nào họ đang tương tác với AI (Nghị viện châu Âu và Hội đồng châu Âu, 2024). Bài học cốt lõi cho Việt Nam: nguyên tắc "trách nhiệm nền tảng" (platform accountability) buộc các công ty mạng xã hội phải chịu trách nhiệm pháp lý về nội

dung trên nền tảng của mình cần được nội luật hóa cho phù hợp.

3.2.2. Trung Quốc, mô hình quản trị không gian mạng chủ quyền với trọng tâm pháp lý AI

Trung Quốc đã xây dựng khung quản lý không gian mạng dựa trên nguyên tắc "chủ quyền không gian mạng" (internet sovereignty), quan điểm rằng mỗi quốc gia có quyền và trách nhiệm quản lý không gian mạng trong phạm vi lãnh thổ theo luật pháp và giá trị của quốc gia mình. "Biện pháp quản lý tạm thời về dịch vụ trí tuệ nhân tạo tạo sinh" ( ) của Trung Quốc, có hiệu lực từ ngày 15/8/2023, là một trong những pháp lý đầu tiên trên thế giới quy định chi tiết về trách nhiệm của các nhà cung cấp dịch vụ AI tạo sinh nội dung: đăng ký và đánh giá trước khi cung cấp dịch vụ ra công chúng; bắt buộc gắn nhãn nội dung do AI tạo ra; từ chối tạo nội dung sai lệch, kích động thù địch hoặc phương hại an ninh quốc gia; lưu trữ nhật ký hoạt động để phục vụ điều tra (Chính phủ Cộng hòa Nhân dân Trung Hoa, 2023). Do khác biệt căn bản về thể chế chính trị, kinh nghiệm của Trung Quốc chỉ có giá trị tham khảo ở khía cạnh kỹ thuật - pháp lý của việc kiểm soát AI tạo sinh nội dung và nhận diện thông tin giả, chứ không phải là khuôn mẫu để áp dụng máy móc.

3.2.3. Singapore, cân bằng tự do ngôn luận và quản trị thông tin trong nền kinh tế mở

Singapore có kinh nghiệm phong phú về quản trị thông tin trong bối cảnh kinh tế mở, đa văn hóa và hội nhập sâu. Đạo luật Bảo vệ khỏi Thông tin Giả mạo và Thao túng Trực tuyến (POFMA, 2019) và Đạo luật An toàn Trực tuyến (2022) (Singapore Ministry of Communications and Information, 2019) là hệ thống pháp lý được thiết kế nhằm cân bằng giữa hai giá trị thường căng thẳng với nhau: bảo vệ tự do ngôn luận và chống thông tin giả. Điểm đáng chú ý của mô hình Singapore là cơ chế "đính chính" thay vì "gỡ bỏ" ngay: khi cơ quan có thẩm quyền xác định một thông tin là sai, họ không xóa nó ngay mà yêu cầu nền tảng gắn nhãn cảnh báo kèm đường dẫn đến thông tin chính thống. Cách tiếp cận kỹ thuật - quản trị này có thể là một gợi ý tham khảo trong việc xử lý thông tin sai lệch mà vẫn hạn chế tác động tiêu cực đến quyền tiếp cận thông tin của người dân.

3.2.4. Hoa Kỳ, kinh nghiệm thực chiến và mô hình hợp tác công - tư

Dù thiếu một luật liên bang thống nhất về thông tin giả, Hoa Kỳ có kinh nghiệm phong phú về đối phó thực tế với các chiến dịch tấn công tư tưởng quy mô lớn sử dụng AI. Trung tâm Gắn kết Toàn cầu (Global Engagement Center - GEC) thuộc Bộ Ngoại giao Mỹ đã công bố nhiều báo cáo về cách thức các thế lực thù địch sử dụng AI trong các chiến dịch gây ảnh hưởng xuyên quốc gia (US Department of State - GEC, 2024). Giá trị tham khảo đáng chú ý từ kinh nghiệm Mỹ là mô hình hợp tác công - tư đa chủ thể: Chính phủ công khai hóa thông tin về các chiến dịch tấn công để báo chí và xã hội cảnh giác; các công ty công nghệ xây dựng công cụ phát hiện video, hình ảnh giả mạo và thông tin giả; các tổ chức kiểm chứng thông tin độc lập đóng vai trò trung gian. UNESCO đúc kết kinh nghiệm toàn cầu trong "Sổ tay Báo chí, Tin giả và Thông tin sai lệch" (2018, tr.9-15), đề xuất mô hình tích hợp bốn trụ cột: giáo dục năng lực truyền thông và thông tin, kiểm chứng thông tin, quản trị nền tảng và đào tạo nhà báo, một khung tham chiếu có giá trị mà Việt Nam có thể nghiên cứu, vận dụng.

3.3. Giá trị tham khảo cho Việt Nam

Cần xác định rõ một nguyên tắc có tính phương pháp luận khi tiếp nhận kinh nghiệm quốc tế: do khác biệt căn bản về thể chế chính trị, kinh nghiệm của các quốc gia, tổ chức nêu trên chỉ có giá trị tham khảo ở khía cạnh nhận diện thông tin giả và các giải pháp kỹ thuật - quản trị, chứ không phải là "bài học" để áp dụng máy móc. Bởi vì, xét đến cùng, bảo vệ nền tảng tư tưởng của Đảng chính là bảo vệ hệ tư tưởng của Đảng, chủ nghĩa Mác - Lênin, tư tưởng Hồ Chí Minh, một nhiệm vụ mang bản chất chính trị, giai cấp sâu sắc, không thể vay mượn từ bên ngoài. Vì vậy, việc tiếp thu kinh nghiệm quốc tế phải hết sức thận trọng, có chọn lọc và phải được "bản địa hóa" cho phù hợp với thể chế chính trị, văn hóa và điều kiện cụ thể của Việt Nam. Trên tinh thần đó, có thể chắt lọc bốn giá trị tham khảo phục vụ trực tiếp cho việc kiến tạo "lá chắn tư tưởng số".

Một là, không có "vắc-xin"đơn lẻ nào có thể giải quyết thách thức này. Kinh nghiệm của EU, Trung Quốc, Singapore và Hoa Kỳ đều cho thấy: chỉ có chiến lược tổng thể, đa tầng, kết hợp pháp lý - kỹ thuật - giáo dục - truyền thông mới có hiệu quả thực sự. Một giải pháp đơn lẻ, dù mạnh đến đâu, đều có thể bị vô hiệu hóa bởi sự tinh vi và linh hoạt không ngừng của AI.

Hai là, tốc độ và tính chủ động là yếu tố quyết định. V.I.Lênin nhiều lần nhấn mạnh rằng trong những thời khắc cách mạng quyết định, sự do dự, chậm trễ là điều nguy hiểm nhất, chủ động nắm thời cơ là yếu tố sống còn. Trong cuộc chiến thông tin trên không gian số, bên nào chủ động trước, cả trong phát triển công cụ kỹ thuật lẫn trong lan tỏa thông tin chính thống sẽ có lợi thế quyết định. Thông tin giả lan truyền theo cấp số nhân trong vài giờ đầu, đây chính là "cửa sổ thời gian vàng" mà lực lượng bảo vệ tư tưởng phải hành động.

Ba là, cần xây dựng đội ngũ chuyên gia liên ngành. Theo tinh thần lời dạy của Chủ tịch Hồ Chí Minh về xây dựng đội ngũ cán bộ "vừa hồng vừa chuyên", trong kỷ nguyên AI, chữ "chuyên" nhất thiết phải bao gồm năng lực số, khả năng nhận diện video, hình ảnh giả mạo, phân tích mạng lưới tài khoản tự động và vận hành hệ thống kiểm chứng thông tin. Không thể đối phó hiệu quả với AI tấn công tư tưởng chỉ bằng đội ngũ cán bộ được đào tạo theo mô hình truyền thống, dù có lý luận vững chắc, nếu thiếu năng lực công nghệ số.

Bốn là, cần tăng cường hợp tác quốc tế trong quản trị AI và bảo vệ tư tưởng số. Tấn công tư tưởng bằng AI không có biên giới quốc gia, các chiến dịch thường có nguồn gốc từ ngoài lãnh thổ, sử dụng hạ tầng đặt ở nhiều quốc gia khác nhau. Hợp tác quốc tế về chia sẻ thông tin, phối hợp pháp lý xuyên biên giới và xây dựng tiêu chuẩn chung về quản trị AI là không thể thiếu, đặc biệt trong khuôn khổ ASEAN.

3.4. Thảo luận

Tổng hợp các kết quả ở mục 3.1, 3.2 và 3.3 dẫn tới một nhận thức có ý nghĩa then chốt đối với toàn bộ chiến lược bảo vệ nền tảng tư tưởng: trọng tâm của cuộc đấu tranh tư tưởng trong kỷ nguyên AI đã dịch chuyển từ bình diện nội dung, nơi các bên cạnh tranh nhau bằng lập luận, lý lẽ và bằng chứng, sang bình diện cơ chế, tức kiến trúc của chính môi trường thông tin mà trong đó nhận thức được hình thành. Ba cơ chế tâm lý - xã hội phân tích ở mục 3.1 không vận hành như những lời "thuyết phục" thông thường; chúng can thiệp vào điều kiện hình thành phán đoán của con người trước khi tư duy lý tính kịp lên tiếng. Nói cách khác, kẻ tấn công không còn chủ yếu tìm cách "thắng trong tranh luận" mà thiết kế sẵn môi trường để người dùng tự đi đến kết luận mà chúng mong muốn. Đây chính là điểm khác biệt về chất so với mọi hình thức

tuyên truyền và chiến tranh tâm lý trước đây và cũng là lời giải đáp cho câu hỏi đặt ra ở phần đầu: vì sao những thủ đoạn ấy lại hiệu quả đến vậy.

Vì sao cơ chế đó có hiệu lực mạnh đến thế, và vì sao AI tạo ra một bước nhảy về chất chứ không đơn thuần về lượng? Câu trả lời nằm ngay trong nền tảng phương pháp luận của bài viết. Khi C.Mác khẳng định "Trong tính hiện thực của nó, bản chất con người là tổng hòa những quan hệ xã hội" (Mác & Ăngghen, 1995, t.3, tr.11), luận điểm ấy đã chỉ ra rằng ý thức của mỗi cá nhân không hình thành trong chân không mà trong mạng lưới quan hệ xã hội bao quanh họ. Trong kỷ nguyên số, một phần ngày càng lớn của mạng lưới ấy được trung gian hóa bởi thuật toán; vì vậy, kẻ kiểm soát thuật toán giành được quyền chi phối một phần môi trường xã hội và qua đó tác động vào chính quá trình hình thành ý thức. Đó là lý do các cơ chế tâm lý không dừng ở việc thuyết phục từng cá nhân riêng lẻ mà có thể tái cấu trúc cả "trường" nhận thức của cộng đồng. Tính mới về chất của AI bắt nguồn từ chỗ nó tháo bỏ giới hạn lao động sống vốn trói buộc mọi chiến dịch tuyên truyền trước đây: lần đầu tiên trong lịch sử, một chủ thể có thể đồng thời thực hiện bốn việc tưởng chừng loại trừ nhau: quy mô đại chúng, cá nhân hóa đến từng người, vận hành tự động và chi phí gần bằng không. Chính tổ hợp "đại chúng + cá nhân hóa + tự động + gần như miễn phí", điều bất khả về mặt lôgíc trong thời đại tiền AI, mới là thứ làm thay đổi bản chất cuộc đấu tranh, đúng như nhận định đã nêu ở Kỳ 1.

Đặt kết quả nghiên cứu trong tương quan với các công trình đi trước càng làm rõ đóng góp riêng của bài viết. Những nghiên cứu kinh điển về thiên kiến nhận thức và "bong bóng lọc", từ Kahneman (2011), Haidt (2012), Asch (1955), Cialdini (2006) đến Pariser (2011), chủ yếu khảo sát các hiện tượng này dưới góc độ tâm lý học ra quyết định, hoặc trong bối cảnh rối loạn thông tin của các nền dân chủ phương Tây và hành vi tiêu dùng. Bài viết kế thừa những phát hiện đó nhưng đặt chúng vào một khung phân tích khác: khung đấu tranh chính trị - tư tưởng có tổ chức, mang bản chất giai cấp. Dưới khung này, những gì các tác giả phương Tây mô tả như "thao túng người tiêu dùng" hay "phân cực xã hội" được nhận diện đúng bản chất là vũ khí trong một cuộc đối đầu hệ tư tưởng, một sự định vị mà bản thân các công trình gốc không đặt ra. So với hướng nghiên cứu phổ biến trong nước về bảo vệ nền tảng tư tưởng, vốn tập trung nhận diện và phản bác các quan điểm sai trái ở bình diện nội dung, bài viết bổ sung một lát cắt còn ít được khai thác: lý giải không chỉ kẻ địch "nói gì" mà cả "hệ

thống chuyển tải khai thác tâm lý con người như thế nào". Ở bình diện kinh nghiệm quốc tế, trong khi nhiều khảo cứu so sánh thường trình bày EU, Trung Quốc, Singapore và Hoa Kỳ như những trường hợp tách rời, mục 3.2 và 3.3 lại rút ra nguyên lý hội tụ chung, đồng thời kiên trì yêu cầu "bản địa hóa" xuất phát từ bản chất chính trị, giai cấp của nhiệm vụ, điều mà các nghiên cứu viết từ lập trường đa nguyên tự do thường không nhấn mạnh.

Về mặt lý luận, kết quả nghiên cứu củng cố và cụ thể hóa luận điểm duy vật biện chứng về quan hệ giữa tồn tại xã hội và ý thức xã hội trong điều kiện mới, đồng thời cho thấy phương pháp luận của chủ nghĩa Mác - Lênin không hề lỗi thời trước AI; trái lại, chính vì không quy hiện tượng về thuần túy kỹ thuật, nó có lợi thế đặc biệt trong việc bóc tách bản chất chính trị ẩn sau lớp vỏ công nghệ. Chỉ dẫn của Chủ tịch Hồ Chí Minh "Biết người cố nhiên là khó, tự biết mình cũng không dễ" (Hồ Chí Minh, 2011, t.5, tr.277) vì thế giữ nguyên giá trị phương pháp: hiểu cơ chế tấn công của đối phương và hiểu chính điểm yếu tâm lý của cộng đồng mình là hai mặt không thể tách rời của một chiến lược phòng thủ đúng đắn. Về mặt thực tiễn, hệ quả quan trọng nhất là: một lá chắn chỉ được thiết kế ở bình diện nội dung, tức chỉ lo phản bác từng luận điệu sai trái, sẽ luôn ở thế bị động và không đủ, bởi nó để nguyên cơ chế khai thác tâm lý tiếp tục vận hành. Do đó, "lá chắn tư tưởng số" tất yếu phải tác động đồng thời trên ba tầng: nội dung, cơ chế và năng lực "miễn dịch" nhận thức của con người, chính định hướng này là cầu nối trực tiếp dẫn sang hệ thống sáu nhóm giải pháp được trình bày ở Kỳ 3. Cũng cần thẳng thắn nhìn nhận giới hạn của nghiên cứu: bài viết chủ yếu dựa trên phương pháp phân tích - tổng hợp tài liệu và so sánh, nên việc đo lường định lượng mức độ tác động thực tế của từng cơ chế trong bối cảnh người dùng Việt Nam vẫn là một hướng cần tiếp tục triển khai bằng khảo sát và thực nghiệm.

4. Kết luận kỳ 2

Phân tích cơ chế tâm lý - xã hội của tấn công tư tưởng bằng AI cho thấy một thực tế quan trọng: kẻ thù không tấn công bằng lý luận mà bằng tâm lý; không cố thuyết phục người dùng mà "thiết kế" môi trường thông tin khiến người dùng "tự thuyết phục bản thân". Đây là sự tinh vi chưa từng thấy trong lịch sử chiến tranh tâm lý. Kinh nghiệm quốc tế cho thấy: đây là vấn đề chung của tất cả các quốc gia trong kỷ nguyên AI, Việt Nam không đơn độc trong cuộc chiến này, và không cần phải "phát minh lại bánh xe"

từ đầu, mà có thể chọn lọc, tiếp thu có phê phán những kinh nghiệm phù hợp với điều kiện và thể chế của mình. Thực tiễn cho thấy, bảo vệ nền tảng tư tưởng trong thời đại số đòi hỏi sự sáng tạo không ngừng, kết hợp chặt chẽ bản lĩnh chính trị với năng lực công nghệ. Đó cũng chính là tinh thần cốt lõi để xây dựng một "lá chắn tư tưởng số" vững chắc, đáp ứng yêu cầu bảo vệ nền tảng tư tưởng của Đảng trong kỷ nguyên phát triển mới theo tinh thần Văn kiện Đại hội đại biểu toàn quốc lần thứ XIV (Đảng Cộng sản Việt Nam, 2026b) và Quy định số 19-QĐ/TW (Đảng Cộng sản Việt Nam, 2026a)./.

Tài liệu tham khảo

1. Asch, S.E. (1955). Opinions and social pressure. Scientific American, 193(5), 31-35.

2. Chính phủ Cộng hòa Nhân dân Trung Hoa. (2023). Biện pháp quản lý tạm thời về dịch vụ trí tuệ nhân tạo tạo sinh ((có hiệu lực từ 15/8/2023).

3. Cialdini, R.B. (2006). Influence: The psychology of persuasion (revised ed.). Harper Business.

4. Đảng Cộng sản Việt Nam. (2026a). Quy định số 19-QĐ/TW ngày 08/4/2026 của Ban Chấp hành Trung ương Đảng về công tác chính trị, tư tưởng trong Đảng.

5. Đảng Cộng sản Việt Nam. (2026b). Văn kiện Đại hội đại biểu toàn quốc lần thứ XIV. Nhà xuất bản Chính trị quốc gia Sự thật.

6. European Commission. (2018). Action plan against disinformation (COM(2018) 794 final).

7. European External Action Service (EEAS). (2024). EUvsDisinfo annual report 2023-2024.

8. Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Pantheon Books.

9. Hồ Chí Minh. (2011). Toàn tập (t.5). Nhà xuất bản Chính trị quốc gia Sự thật.

10. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

11. Mác, C. & Ăngghen, Ph. (1995). Toàn tập (t.3). Nhà xuất bản Chính trị quốc gia Sự thật.

12. Nghị viện châu Âu và Hội đồng châu Âu. (2024). Đạo luật Trí tuệ nhân tạo (Quy định (EU) 2024/1689) . Công báo Liên minh châu Âu.

13. Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin Press.

14. Pew Research Center. (2024). News consumption across social media in 2024.

15. Reuters Institute for the Study of Journalism. (2024). Digital news report 2024. University of Oxford.

16. Singapore Ministry of Communications and Information. (2019). Protection from Online Falsehoods and Manipulation Act (POFMA) 2019; Online Safety (Miscellaneous Amendments) Act 2022.

17. UNESCO. (2018). Journalism, "fake news" and disinformation: A handbook for journalism education and training.

18. US Department of State - Global Engagement Center. (2024). Report on global influence operations using AI.