

AI TẠO SINH TRONG HOẠT ĐỘNG XUẤT BẢN: THÁCH THỨC AN TOÀN THÔNG TIN VÀ GIẢI PHÁP QUẢN TRỊ NỘI DUNG

TRẦN THỌ XƯƠNG
NXB Giáo dục Việt Nam
TRẦN THU QUỲNH
VŨ THỊ NGỌC THÙY
TRẦN THỊ MAI DUNG
Học viện Báo chí và Tuyên truyền

Nhận bài ngày 20/7/2026. Sửa chữa xong 20/8/2026. Duyệt đăng 03/9/2026.

Abstract

Generative artificial intelligence (GenAI) is transforming publishing while presenting challenges to information security and content quality. Through a literature review and the analysis and synthesis of relevant sources, this study identifies key risks: misinformation, fabricated content and deepfakes, algorithmic bias, data leakage, intellectual property infringement, and a diminishing role for humans in information verification. Drawing on these findings, the study proposes content governance measures to safeguard information security and maintain human oversight in digital publishing.

Keywords: Content governance, digital publishing, generative AI, information security, information verification.

1. Đặt vấn đề

Sự phát triển nhanh chóng của trí tuệ nhân tạo tạo sinh (GenAI) đang tạo ra những thay đổi sâu sắc trong nhiều lĩnh vực. Với khả năng tạo lập văn bản, hình ảnh, âm thanh và video, GenAI ngày càng tác động mạnh mẽ đến sáng tạo nội dung, giáo dục, nghiên cứu khoa học và truyền thông (Lorenz et al., 2023). Trong lĩnh vực xuất bản, AI hỗ trợ các khâu sáng tác, biên tập, dịch thuật, thiết kế, tiếp thị và góp phần tái cấu trúc quy trình nghiệp vụ của các nhà xuất bản (Ryzhko et al., 2024). GenAI mang lại nhiều cơ hội cho ngành xuất bản như nâng cao năng suất, tối ưu hóa quy trình sản xuất nội dung và cá nhân hóa trải nghiệm độc giả (Olszewski & Krawczyk, 2026; Saikaly, 2023). Tuy nhiên, công nghệ này cũng đặt ra những thách thức đối với an toàn thông tin và chất lượng nội dung, đặc biệt là thông tin sai lệch, nội dung giả mạo và deepfake. Hiện tượng “ảo giác AI” (AI hallucination) có thể tạo ra dữ liệu, tài liệu tham khảo hoặc thông tin không có thật nhưng được trình bày dưới hình thức thuyết phục, làm gia tăng nguy cơ sai lệch trong hoạt động xuất bản (Giraldo-Barreto, 2026). Bên cạnh đó, việc sử dụng AI còn đặt ra những vấn đề về bảo mật dữ liệu và quyền sở hữu trí tuệ, đặc biệt khi các mô hình được huấn luyện trên khối lượng lớn dữ liệu có bản quyền (U.S. Copyright Office, 2025). Từ thực tiễn trên, bài viết phân tích những thách thức đối với an toàn thông tin trong hoạt động xuất bản dưới tác động của AI tạo sinh, tập trung vào thông tin sai lệch, giả mạo nội dung, rò rỉ dữ liệu, xâm phạm quyền sở hữu trí tuệ và suy giảm vai trò kiểm chứng của con người. Trên cơ sở đó, bài viết đề xuất các giải pháp quản trị nội dung nhằm khai thác hiệu quả lợi ích của AI và giảm thiểu rủi ro an toàn thông tin trong môi trường xuất bản số.

2. Phương pháp nghiên cứu

2.1. Phương pháp nghiên cứu tài liệu

Phương pháp này được sử dụng để thu thập, hệ thống hóa và đánh giá các quan điểm lý luận, chính sách và kết quả nghiên cứu liên quan đến GenAI trong hoạt động xuất bản. Nguồn tài liệu gồm các

Email: xuongxb.gd@gmail.com

khuyến nghị, báo cáo và nghiên cứu của UNESCO, OECD, Văn phòng Bản quyền Hoa Kỳ, Nghị viện châu Âu; các nghiên cứu quốc tế về AI trong ngành xuất bản; cùng các văn bản của Đảng, chính sách, pháp luật của Nhà nước và tài liệu chuyên ngành tại Việt Nam. Các tài liệu được tập trung khai thác nhằm làm rõ sự tác động của GenAI đến hoạt động xuất bản và những vấn đề liên quan đến an toàn thông tin, chất lượng nội dung, bảo mật dữ liệu và quyền sở hữu trí tuệ.

2.2. Phương pháp phân tích - tổng hợp

Phương pháp phân tích - tổng hợp được sử dụng để nhận diện và hệ thống hóa các nhóm rủi ro an toàn thông tin phát sinh trong hoạt động xuất bản dưới tác động của GenAI. Trên cơ sở phân tích các kết quả nghiên cứu, khuyến nghị chính sách và kinh nghiệm quốc tế, bài viết tổng hợp những vấn đề chủ yếu về tính chính xác và độ tin cậy của thông tin, bảo mật dữ liệu, quyền sở hữu trí tuệ và vai trò kiểm chứng của con người. Kết quả phân tích là cơ sở để đề xuất các giải pháp quản trị nội dung phù hợp với bối cảnh chuyển đổi số của ngành xuất bản ở Việt Nam.

3. Nội dung nghiên cứu

3.1. GenAI và sự thay đổi trong hoạt động xuất bản

Sự phát triển của GenAI đang tạo ra những thay đổi đáng kể đối với hoạt động xuất bản trên phạm vi toàn cầu. Khác với các công nghệ hỗ trợ truyền thống, GenAI có khả năng tạo lập nội dung mới dưới nhiều hình thức như văn bản, hình ảnh, âm thanh và video, qua đó tác động trực tiếp đến nhiều công đoạn trong chuỗi giá trị xuất bản. GenAI đang trở thành một công nghệ nền tảng có ảnh hưởng sâu rộng đến hoạt động sáng tạo nội dung, giáo dục, nghiên cứu và truyền thông (Lorenz et al., 2023). Trong lĩnh vực xuất bản, AI được ứng dụng ngày càng phổ biến trong hoạt động sáng tác và biên tập. Các công cụ như ChatGPT, Claude hay Gemini có khả năng hỗ trợ xây dựng đề cương, gợi ý ý tưởng, tóm tắt tài liệu, hiệu đính ngôn ngữ và chuẩn hóa văn phong. GenAI đang góp phần tái cấu trúc các quy trình nghiệp vụ của ngành xuất bản, đặc biệt ở các khâu xây dựng nội dung và biên tập bản thảo (Ryzhko et al., 2024). Đồng thời, AI đã trở thành công cụ hỗ trợ quan trọng đối với tác giả, biên tập viên và các nhà xuất bản trong quá trình sản xuất nội dung (Giraldo-Barreto, 2026). Bên cạnh sáng tác và biên tập, AI còn tạo ra những thay đổi đáng kể trong hoạt động dịch thuật, thiết kế, marketing và phát hành xuất bản phẩm. Các hệ thống AI có khả năng dịch văn bản nhanh với chi phí thấp hơn so với phương pháp truyền thống, góp phần mở rộng khả năng tiếp cận tri thức của độc giả ở nhiều quốc gia (Saikaly, 2023). Trong lĩnh vực thiết kế, các công cụ tạo sinh hình ảnh hỗ trợ tạo minh họa, bìa sách và các sản phẩm truyền thông trực quan, qua đó góp phần rút ngắn thời gian sản xuất và giảm chi phí thiết kế. Trong hoạt động marketing và phát hành, AI được sử dụng để phân tích hành vi người dùng, dự báo nhu cầu thị trường, tối ưu hóa hoạt động quảng bá và đề xuất nội dung phù hợp với từng nhóm độc giả. Nhiều nhà xuất bản lớn trên thế giới đã ứng dụng AI để cá nhân hóa trải nghiệm đọc, dự báo doanh số và nâng cao hiệu quả kinh doanh (Saikaly, 2023). AI cũng góp phần hình thành hệ sinh thái xuất bản dựa trên dữ liệu, trong đó nội dung được phân phối và tiếp cận độc giả theo hướng ngày càng cá nhân hóa (Olszewski & Krawczyk, 2026). Những thay đổi trên cho thấy GenAI đang trở thành một thành tố quan trọng trong hệ sinh thái xuất bản số, góp phần nâng cao năng suất, giảm chi phí và mở rộng khả năng tiếp cận tri thức. Tuy nhiên, việc ứng dụng AI cũng đặt ra những thách thức liên quan đến an toàn thông tin, tính xác thực của nội dung, quyền sở hữu trí tuệ và trách nhiệm nghề nghiệp trong hoạt động xuất bản.

3.2. Các rủi ro an toàn thông tin trong hoạt động xuất bản dưới tác động của GenAI

3.2.1. Nguy cơ phát tán thông tin sai lệch

Một trong những rủi ro lớn của GenAI đối với hoạt động xuất bản là khả năng tạo ra nội dung sai lệch nhưng được trình bày dưới hình thức hợp lý và thuyết phục. Các mô hình ngôn ngữ lớn hoạt động dựa trên cơ chế dự đoán xác suất từ ngữ tiếp theo từ dữ liệu huấn luyện, thay vì khả năng nhận thức và kiểm chứng sự thật như con người (Giraldo-Barreto, 2026). Do đó, AI có thể tạo ra thông tin không chính xác, dữ liệu không tồn tại hoặc tài liệu tham khảo giả nhưng vẫn được trình bày dưới hình thức học thuật, thường được gọi là hiện tượng “ảo giác AI” (AI hallucination). Trong hoạt động xuất bản, đặc biệt là xuất bản học thuật và khoa học, nội dung sai lệch do AI tạo ra có thể ảnh hưởng đến chất

lượng xuất bản phẩm, làm suy giảm độ tin cậy của thông tin và gây khó khăn cho quá trình kiểm chứng nguồn tin. Đồng thời, khả năng sản xuất nội dung với tốc độ lớn của AI có thể làm gia tăng nguy cơ lan truyền thông tin không được kiểm chứng trong môi trường số.

3.2.2. Nguy cơ giả mạo nội dung và deepfake

Sự phát triển của GenAI đã mở rộng từ văn bản sang hình ảnh, âm thanh và video, cho phép tạo ra các nội dung giả với mức độ chân thực ngày càng cao, làm gia tăng nguy cơ deepfake trong lĩnh vực truyền thông và xuất bản. Trong hoạt động xuất bản, AI có thể bị lợi dụng để tạo hình ảnh minh họa, tài liệu giả hoặc nội dung mang danh nghĩa của các tác giả, nhà nghiên cứu và tổ chức có uy tín. Thực tế đã xuất hiện những cuốn sách do AI tạo ra nhưng được gắn tên của các tác giả nổi tiếng mà không có sự cho phép của họ (Giraldo-Barreto, 2026). Điều này không chỉ xâm phạm quyền tác giả mà còn đe dọa tính xác thực của hệ sinh thái xuất bản số. Bên cạnh nguy cơ giả mạo, deepfake còn có thể ảnh hưởng đến niềm tin và khả năng đánh giá thông tin của công chúng. Việc thường xuyên tiếp xúc với deepfake và thông tin không kiểm chứng có thể dẫn đến lo âu thông tin, gia tăng tải nhận thức và làm suy giảm khả năng đánh giá thông tin của người dùng (Nguyễn Đức Quang, Phan Thị Thanh Chi, 2025). Trong hoạt động xuất bản, những hệ quả này có thể ảnh hưởng đến khả năng tiếp nhận tri thức của độc giả và uy tín của các cơ quan xuất bản trong môi trường số.

3.2.3. Nguy cơ thiên kiến thuật toán và phân biệt đối xử trong nội dung

GenAI có nguy cơ phản ánh và khuếch đại những thiên kiến, định kiến hoặc khuôn mẫu tồn tại trong dữ liệu huấn luyện, khiến nội dung được tạo ra có thể thiếu tính khách quan, cân bằng hoặc chứa đựng biểu hiện phân biệt đối xử. Do đó, các hệ thống AI cần được kiểm soát nhằm hạn chế tác động phân biệt đối xử và giảm thiểu sự khuếch đại thiên kiến bởi thuật toán (UNESCO, 2021). Trong hoạt động xuất bản, nếu nội dung do AI tạo ra thiếu sự thẩm định của tác giả và biên tập viên, những thiên kiến này có thể được đưa vào bản thảo và lan truyền tới độc giả. Vì vậy, việc kiểm soát nội dung không chỉ dừng ở xác minh tính chính xác và xác thực mà cần đánh giá cả tính khách quan, cân bằng và công bằng, đồng thời duy trì vai trò thẩm định của con người trong quy trình biên tập.

3.2.4. Rủi ro đối với dữ liệu và bản thảo xuất bản

Việc sử dụng các nền tảng AI công cộng trong sáng tác và biên tập đặt ra những thách thức đối với bảo mật dữ liệu, đặc biệt khi bản thảo, dữ liệu nghiên cứu hoặc nội dung chưa công bố được tải lên để xử lý, hiệu đính hoặc tóm tắt. Nhiều tổ chức học thuật quốc tế như COPE và WAME đã khuyến cáo không tải bản thảo chưa công bố, nội dung phản biện kín hoặc dữ liệu nghiên cứu nhạy cảm lên các nền tảng AI công cộng do nguy cơ rò rỉ thông tin và mất quyền kiểm soát dữ liệu (Đào Trung Hiếu, 2025). Đối với các nhà xuất bản, việc chia sẻ dữ liệu và bản thảo chưa công bố với các hệ thống AI bên ngoài có thể làm gia tăng nguy cơ rò rỉ và mất quyền kiểm soát thông tin. Đây không chỉ là vấn đề bảo mật dữ liệu mà còn liên quan đến quyền sở hữu trí tuệ, bí mật kinh doanh và uy tín nghề nghiệp.

3.2.5. Thách thức về bản quyền và sở hữu trí tuệ

Sự phát triển của GenAI đặt ra nhiều vấn đề pháp lý và đạo đức về sở hữu trí tuệ, đặc biệt khi các mô hình AI được huấn luyện trên khối lượng lớn dữ liệu có chứa các tác phẩm được bảo hộ bản quyền. Việc sử dụng các tác phẩm này để huấn luyện AI gây nhiều tranh luận do liên quan trực tiếp đến quyền sao chép và khai thác tác phẩm của chủ sở hữu quyền (U.S. Copyright Office, 2025). Hai vấn đề pháp lý đáng chú ý khác là việc sử dụng dữ liệu có bản quyền để huấn luyện AI và khả năng bảo hộ đối với nội dung do AI tạo ra; nếu quyền lợi của người sáng tạo không được bảo đảm, động lực sáng tạo và nguồn dữ liệu chất lượng cao cho phát triển AI có thể bị ảnh hưởng (Yang & Zhang, 2025). Bên cạnh đó, việc xác định chủ thể quyền đối với sản phẩm được tạo ra với sự hỗ trợ của AI vẫn là thách thức đối với các hệ thống pháp luật hiện hành (Lucchi, 2025). Trong lĩnh vực xuất bản, vấn đề này có thể dẫn đến tranh chấp về quyền sở hữu nội dung, quyền khai thác thương mại và trách nhiệm đối với hành vi vi phạm bản quyền.

3.2.6. Sự suy giảm vai trò kiểm chứng của con người

Việc phụ thuộc quá mức vào AI trong sáng tác, biên tập và đánh giá nội dung có thể làm suy giảm vai trò kiểm chứng và trách nhiệm nghề nghiệp của con người, đặc biệt khi kết quả do AI tạo ra được sử

dụng mà không qua kiểm tra, xác minh. Ngay cả các công cụ phát hiện nội dung AI cũng chưa đạt độ chính xác tuyệt đối; OpenAI từng ngừng triển khai AI Classifier và Turnitin khuyến cáo không sử dụng kết quả phát hiện AI làm bằng chứng duy nhất trong các quyết định học thuật (Đào Trung Hiếu, 2025). Điều này cho thấy việc đánh giá cuối cùng vẫn cần có sự tham gia của con người. Trong xuất bản học thuật đã xuất hiện những trường hợp bài báo chứa nội dung do AI tạo ra nhưng không được phát hiện trong quá trình biên tập và phản biện (Giraldo-Barreto, 2026). Vì vậy, sự suy giảm vai trò của tác giả, biên tập viên và phản biện viên có thể ảnh hưởng đến chất lượng và độ tin cậy của xuất bản phẩm. Việc ứng dụng AI cần bảo đảm trách nhiệm giải trình của con người, trong đó con người giữ vai trò trung tâm trong giám sát, đánh giá và ra quyết định đối với nội dung có sự hỗ trợ của AI (UNESCO, 2021).

3.2.7. Thực trạng ứng dụng GenAI tại các nhà xuất bản ở Việt Nam

Tại Việt Nam, GenAI đang tác động ngày càng rõ nét đến hoạt động xuất bản và thúc đẩy các đơn vị trong ngành ứng dụng công nghệ vào quy trình sản xuất nội dung. AI đã được tiếp cận và thử nghiệm ở nhiều công đoạn như biên soạn, dịch thuật, hiệu đính, thiết kế, sản xuất sách nói và truyền thông xuất bản phẩm. Trong lĩnh vực sách nói, công nghệ chuyển đổi văn bản thành giọng nói (Text-to-Speech) kết hợp với AI mở ra khả năng tạo giọng đọc tự động, góp phần rút ngắn thời gian sản xuất. Trong thiết kế và minh họa, các công cụ GenAI hình ảnh như Midjourney và DALL-E cũng có thể hỗ trợ phác thảo ý tưởng, thiết kế bìa và xây dựng hình ảnh minh họa. Những ứng dụng này cho thấy AI đang mở rộng khả năng hỗ trợ các khâu của hoạt động xuất bản và đặt ra yêu cầu thích ứng với sự phát triển của công nghệ. Tuy nhiên, việc ứng dụng GenAI cũng làm xuất hiện những vấn đề về quản trị nội dung và an toàn thông tin. Sử dụng AI để biên soạn, dịch thuật hoặc hiệu đính mà thiếu kiểm chứng có thể làm gia tăng nguy cơ thông tin sai lệch, “ảo giác AI” hoặc nội dung chưa phù hợp với bối cảnh văn hóa – xã hội. Hình ảnh do AI tạo ra nếu không được kiểm tra cũng có thể xuất hiện các lỗi ảnh hưởng đến chất lượng xuất bản phẩm. Đặc biệt, việc đưa bản thảo, dữ liệu nghiên cứu hoặc nội dung chưa công bố lên các nền tảng AI công cộng có thể dẫn đến nguy cơ rò rỉ thông tin, mất quyền kiểm soát dữ liệu và phát sinh vấn đề về quyền sở hữu trí tuệ.

Thực trạng trên đặt ra yêu cầu các cơ quan xuất bản Việt Nam xây dựng quy định và quy trình sử dụng AI phù hợp, xác định rõ trách nhiệm của người sử dụng và duy trì vai trò của tác giả, biên tập viên trong kiểm tra, thẩm định nội dung. Đây là cơ sở để khai thác lợi ích của GenAI, đồng thời hạn chế rủi ro đối với chất lượng nội dung và an toàn thông tin.

3.3. Một số kinh nghiệm quốc tế trong quản trị AI và an toàn thông tin

Trước những cơ hội và rủi ro của GenAI, nhiều tổ chức quốc tế và chủ thể trong lĩnh vực xuất bản đã xây dựng các nguyên tắc và cơ chế nhằm bảo đảm việc ứng dụng AI an toàn, minh bạch và có trách nhiệm. Từ kinh nghiệm quốc tế, có thể khái quát một số định hướng chính có giá trị tham khảo đối với hoạt động xuất bản: 1) Quản trị AI cần dựa trên các nguyên tắc đạo đức, minh bạch, trách nhiệm giải trình và sự giám sát của con người. UNESCO nhấn mạnh việc phát triển và ứng dụng AI phải gắn với minh bạch, trách nhiệm giải trình, bảo vệ quyền con người và duy trì sự giám sát của con người (UNESCO, 2021). Đối với hoạt động xuất bản, những nguyên tắc này là cơ sở để xác định trách nhiệm của các chủ thể khi sử dụng AI và góp phần bảo đảm độ tin cậy của nội dung; 2) Quản trị AI cần dựa trên mức độ rủi ro, đồng thời tăng cường tính minh bạch và khả năng nhận diện nội dung do AI tạo ra. Sự phát triển của GenAI đòi hỏi các cơ chế kiểm soát phù hợp nhằm hạn chế rủi ro đối với quyền sở hữu trí tuệ, quyền riêng tư và độ tin cậy của thông tin (Lucchi, 2025). Cách tiếp cận này được thể hiện trong các chính sách quản trị AI của Liên minh châu Âu, trong đó chú trọng phân loại rủi ro và nghĩa vụ minh bạch đối với nội dung do AI tạo ra; 3) Mô hình quản trị AI cần linh hoạt và phù hợp với hệ thống pháp luật cũng như đặc điểm của từng hệ sinh thái sáng tạo. Không tồn tại một mô hình quản trị bản quyền AI cố định mà cách thức quản trị cần được điều chỉnh theo đặc điểm của từng hệ sinh thái sáng tạo (Yang & Zhang, 2025). Thực tiễn tại Hoa Kỳ, Liên minh châu Âu và Trung Quốc cho thấy các quốc gia, khu vực có thể lựa chọn những cách tiếp cận khác nhau, nhưng đều hướng đến kiểm soát rủi ro, bảo vệ quyền sở hữu trí tuệ và tăng cường trách nhiệm đối với việc sử dụng AI; 4) Các nguyên tắc quản trị cần được cụ thể hóa thành quy định và quy trình sử dụng AI trong hoạt động xuất bản. AI đang được ứng dụng trong các hoạt động như biên tập, dịch thuật, phát hiện đạo văn, tóm tắt nội dung và hỗ trợ phản biện

học thuật (Saikaly, 2023). Cùng với quá trình này, các tổ chức xuất bản cần xác định rõ trách nhiệm của tác giả và biên tập viên, bảo mật bản thảo, minh bạch việc sử dụng AI và duy trì sự kiểm soát của con người đối với nội dung trước khi xuất bản.

Nhìn chung, kinh nghiệm quốc tế cho thấy quản trị AI trong xuất bản cần kết hợp bốn yếu tố: nền tảng đạo đức và trách nhiệm của con người; quản trị theo rủi ro và minh bạch nội dung; mô hình quản trị phù hợp với từng hệ sinh thái; và cụ thể hóa các nguyên tắc thành quy trình nghiệp vụ. Đây là những kinh nghiệm có giá trị tham khảo đối với các cơ quan xuất bản Việt Nam trong xây dựng cơ chế quản trị nội dung và bảo đảm an toàn thông tin.

3.4. Giải pháp quản trị nội dung nhằm bảo đảm an toàn thông tin trong hoạt động xuất bản

3.4.1. Hoàn thiện quy trình kiểm chứng nội dung

Trong bối cảnh GenAI có thể tạo ra thông tin sai lệch, dữ liệu không có thật hoặc tài liệu tham khảo giả nhưng được trình bày dưới hình thức thuyết phục (Giraldo-Barreto, 2026), các cơ quan xuất bản cần tăng cường quy trình kiểm chứng nội dung. Những thông tin quan trọng như dữ liệu khoa học, số liệu thống kê và tài liệu tham khảo cần được kiểm chứng từ nhiều nguồn; đối với xuất bản phẩm có tính học thuật cao cần kết hợp biên tập với thẩm định chuyên gia để bảo đảm tính chính xác và độ tin cậy. Kiểm chứng cần trở thành bước bắt buộc trong quy trình biên tập, áp dụng không chỉ với văn bản mà cả hình ảnh, biểu đồ, dữ liệu và sản phẩm đa phương tiện có sự hỗ trợ của AI. Việc ứng dụng AI phải bảo đảm tính minh bạch và trách nhiệm giải trình của con người (UNESCO, 2021). Vì vậy, AI chỉ nên đóng vai trò hỗ trợ, còn trách nhiệm cuối cùng về chất lượng, tính chính xác và tính hợp pháp của xuất bản phẩm thuộc về tác giả, biên tập viên và cơ quan xuất bản.

3.4.2. Xây dựng chính sách sử dụng AI trong nhà xuất bản

Các nhà xuất bản cần xây dựng chính sách nội bộ về sử dụng AI nhằm bảo đảm tính minh bạch, trách nhiệm giải trình và kiểm soát rủi ro. Chính sách cần quy định rõ phạm vi sử dụng AI trong các khâu sáng tác, biên tập, dịch thuật, thiết kế và truyền thông; đồng thời xác định những công việc AI chỉ được hỗ trợ mà không thay thế hoàn toàn vai trò của con người. Bên cạnh đó, cần xác định trách nhiệm của tác giả, biên tập viên và các chủ thể liên quan đối với nội dung được tạo ra hoặc hỗ trợ bởi AI. Việc ứng dụng AI phải gắn với nguyên tắc minh bạch và trách nhiệm giải trình của con người (UNESCO, 2021). Các nhà xuất bản có thể tham khảo mô hình “đèn giao thông”, phân loại hoạt động sử dụng AI thành ba nhóm: được phép, hạn chế và cấm (Đào Trung Hiếu, 2025). Theo đó, AI có thể được sử dụng trong một số hoạt động hỗ trợ; việc sử dụng AI để biên soạn nội dung chuyên môn cần được kiểm soát chặt chẽ; còn các hành vi như tạo dữ liệu giả, nguồn tham khảo không có thật hoặc che giấu việc sử dụng AI cần bị nghiêm cấm. Đồng thời, cần có cơ chế kê khai, kiểm tra và phê duyệt việc sử dụng AI trước khi công bố nội dung.

3.4.3. Tăng cường bảo mật dữ liệu xuất bản

Trong môi trường xuất bản số, các bản thảo chưa công bố, dữ liệu nghiên cứu, hồ sơ tác giả, dữ liệu độc giả và tài liệu có giá trị thương mại là những tài sản thông tin cần được bảo vệ. Việc sử dụng các nền tảng AI công cộng để hỗ trợ viết, biên tập và xử lý nội dung có thể làm gia tăng nguy cơ dữ liệu bị lưu trữ, phân tích hoặc sử dụng ngoài sự kiểm soát của chủ sở hữu. Vì vậy, các nhà xuất bản cần quản lý dữ liệu trong toàn bộ vòng đời thông tin, từ hình thành bản thảo, biên tập, lưu trữ, trao đổi đến phát hành; đồng thời phân quyền truy cập phù hợp với chức năng, nhiệm vụ của từng cá nhân. Các biện pháp quản trị cần kết hợp với giải pháp kỹ thuật như mã hóa dữ liệu, xác thực người dùng, sao lưu định kỳ và giám sát truy cập. Đặc biệt, cần quy định rõ việc sử dụng AI đối với bản thảo chưa công bố và dữ liệu nhạy cảm, chỉ cho phép chia sẻ với các hệ thống AI khi đáp ứng yêu cầu về an toàn và kiểm soát dữ liệu.

3.4.4. Nâng cao năng lực AI và an toàn thông tin cho biên tập viên

Sự phát triển của GenAI đang đặt ra những yêu cầu mới đối với năng lực của biên tập viên. Bên cạnh các nghiệp vụ truyền thống, biên tập viên cần có khả năng nhận diện nội dung do AI tạo ra, phát hiện thông tin sai lệch, kiểm chứng nguồn tin, đánh giá nguy cơ vi phạm bản quyền và bảo đảm an toàn thông tin. Cho nên, năng lực biên tập cần được bổ sung các kiến thức và kỹ năng về AI, quản trị thông tin và tư duy phản biện. Các cơ quan xuất bản cần tăng cường đào tạo cho biên tập viên về nhận diện “ảo giác AI”, kiểm chứng và đối chiếu nguồn tin, phát hiện nội dung giả mạo, bảo vệ dữ liệu và quyền sở hữu trí tuệ.

Đồng thời, biên tập viên cần được trang bị kỹ năng sử dụng AI có trách nhiệm, khai thác hiệu quả công nghệ nhưng không phụ thuộc vào kết quả do AI tạo ra. Việc nâng cao các năng lực này góp phần duy trì vai trò kiểm chứng của con người và bảo đảm chất lượng, độ tin cậy của nội dung xuất bản.

3.4.5. Kết hợp AI với cơ chế kiểm duyệt của con người

Mặc dù GenAI có thể hỗ trợ xử lý dữ liệu, phát hiện lỗi, tóm tắt, dịch thuật và đề xuất chỉnh sửa, công nghệ này không thể chịu trách nhiệm đối với tính chính xác, tính hợp pháp và hệ quả của nội dung được tạo ra. Những hạn chế như “ảo giác AI”, thông tin sai lệch, trích dẫn giả và nguy cơ vi phạm bản quyền cho thấy AI chưa thể thay thế vai trò đánh giá và phán đoán của con người trong hoạt động xuất bản. Do đó, việc ứng dụng AI cần gắn với cơ chế giám sát và trách nhiệm giải trình của con người. Mô hình “Human-in-the-loop” là hướng tiếp cận phù hợp, trong đó AI đảm nhận các công việc hỗ trợ như xử lý dữ liệu, rà soát lỗi kỹ thuật, tóm tắt hoặc đề xuất chỉnh sửa; còn tác giả, biên tập viên và các chủ thể chuyên môn thực hiện kiểm tra, đánh giá và đưa ra quyết định cuối cùng. Sự kết hợp giữa khả năng xử lý của AI với tư duy phản biện và kinh nghiệm chuyên môn của con người góp phần nâng cao hiệu quả, đồng thời bảo đảm chất lượng và độ tin cậy của nội dung xuất bản.

4. Kết luận

Sự phát triển nhanh chóng của GenAI đang tạo ra những thay đổi sâu sắc trong hoạt động xuất bản, góp phần nâng cao năng suất, tối ưu hóa quy trình sản xuất nội dung và thúc đẩy chuyển đổi số của ngành. Tuy nhiên, kết quả nghiên cứu cho thấy việc ứng dụng GenAI cũng làm gia tăng nhiều rủi ro đối với an toàn thông tin và chất lượng nội dung, nổi bật là phát tán thông tin sai lệch, giả mạo nội dung và deepfake, thiên kiến thuật toán, rò rỉ dữ liệu bản thảo, xâm phạm quyền sở hữu trí tuệ và suy giảm vai trò kiểm chứng của con người. Những rủi ro này không chỉ ảnh hưởng đến tính chính xác, độ tin cậy của xuất bản phẩm mà còn đặt ra thách thức đối với uy tín và trách nhiệm của các cơ quan xuất bản trong môi trường số. Vì vậy, để khai thác hiệu quả tiềm năng của GenAI đồng thời kiểm soát các rủi ro phát sinh, các cơ quan xuất bản cần xây dựng cơ chế quản trị nội dung phù hợp, trong đó chú trọng hoàn thiện quy trình kiểm chứng thông tin, xây dựng chính sách sử dụng AI, tăng cường bảo mật dữ liệu, nâng cao năng lực AI và an toàn thông tin cho đội ngũ biên tập viên, đồng thời duy trì cơ chế giám sát của con người theo mô hình “Human-in-the-loop”. Việc ứng dụng AI cần gắn với các nguyên tắc minh bạch, trách nhiệm giải trình và kiểm soát của con người, qua đó hình thành mô hình phối hợp hiệu quả giữa công nghệ và năng lực biên tập chuyên nghiệp, góp phần nâng cao chất lượng nội dung và bảo đảm an toàn thông tin trong hệ sinh thái xuất bản số.

Tài liệu tham khảo

- Đào Trung Hiếu (2025). Sử dụng trí tuệ nhân tạo trong nghiên cứu khoa học: Cơ hội, thách thức và vấn đề liên chính học thuật. *Tạp chí Khoa học Trường Đại học Mở Hà Nội*, 95, 89-96. <https://doi.org/10.59266/houjs.2025.667>.
- Giraldo-Barreto, M. (2026). Generative artificial intelligence in the publishing industry: Adoption, use, intellectual property, and other challenges. *Frontiers in Research Metrics and Analytics*, 11, 1759242. <https://doi.org/10.3389/frma.2026.1759242>.
- Lorenz, P., Perset, K., & Berryhill, J. (2023). *Initial policy considerations for generative artificial intelligence* (OECD Artificial Intelligence) Papers No. 1. OECD Publishing. <https://doi.org/10.1787/fae2d1e6-en>.
- Lucchi, N. (2025). *Generative AI and copyright: Training, creation, regulation* (Study PE 774.095). European Parliament, Policy Department for Justice, Civil Liberties and Institutional Affairs. <https://doi.org/10.2861/0365517>.
- Olszewski, R., & Krawczyk, D. (2026). Impact of artificial intelligence on the book industry: A path toward the knowledge economy. *Knowledge Economy and Lifelong Learning*, 2(1), 67-98. [https://doi.org/10.61093/kell.2\(1\).67-98.2026](https://doi.org/10.61093/kell.2(1).67-98.2026).
- Ryzhko, O., Krainikova, T., Vodolazka, S., & Sokolova, K. (2024). Generative AI changes the book publishing industry: Reengineering of business processes. *Communication & Society*, 37(3), 255-271. <https://doi.org/10.15581/003.37.3.255-271>.
- Saikaly, T. (2023, November 8). *The transformative role of artificial intelligence in the publishing industry*. FADEL.
- UNESCO. (2022). *Recommendation on the ethics of artificial intelligence*. UNESCO.
- U.S. Copyright Office. (2025). *Copyright and artificial intelligence, part 3: Generative AI training* (Pre-publication version). Library of Congress.
- Yang, S. A., & Zhang, A. H. (2024). *Generative AI and copyright: A dynamic perspective*. arXiv. <https://arxiv.org/abs/2402.17801>.
- Saikaly, T. (2023). *The transformative role of artificial intelligence in the publishing industry*. FADEL.
- UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO.
- U.S. Copyright Office. (2025). *Copyright and artificial intelligence: Part 3 - Generative AI training*. Library of Congress.
- Yang, S. A., & Zhang, A. H. (2025). *Generative AI and copyright: A dynamic perspective*. <https://doi.org/10.48550/arXiv.2402.17801>.