Editorial Control of AI in Investigative Journalism

Kostis Spyroglou

Investigative Journalism

[Professor name redacted]

16 October 2026

Word count: 2,844

Executive Summary

AI can help investigative newsrooms examine material beyond their existing resources, but it also shapes which evidence reporters encounter and which interpretations they pursue. This paper asks how editors-in-chief and newsroom management in Europe can preserve editorial control while using that new investigative capacity. Drawing on particular investigative cases and newsroom research, the paper identifies two requirements: AI literacy to evaluate systems and organisational support to act on those evaluations. It compares five responses and recommends a designated AI specialist along with a written newsroom policy. The specialist would evaluate applications and support colleagues; the policy would specify authorisation, verification, confidentiality, and publication responsibilities. Smaller outlets could allocate an existing role or share expertise. A twelve-month pilot should evaluate total task time, errors, staff competence, and the quality of oversight before further expansion.

Editorial Control of AI in Investigative Journalism

Introduction

Artificial intelligence is expanding the scale and scope of investigative journalism. Beyond accelerating familiar tasks, it lets journalists search large collections of documents, connect dispersed records, and identify patterns that might otherwise have been beyond their available resources (Center for News, Technology & Innovation [CNTI], 2026). OpenAI (2025), for example, markets a tool that analyses hundreds of online sources within minutes. In open-source investigations, Bellingcat has used machine learning to identify social-media posts potentially documenting civilian harm, and BBC investigators have used language models to prioritise material on the war in Ukraine (Online News Association, 2024; Ramalho & Waters, 2026). The International Consortium of Investigative Journalists, with Oslo Metropolitan University and the broadcaster NRK, developed a machine-learning tool that identified around 500 passport scans among more than 110,000 leaked documents, helping trace those behind offshore companies (Doumouro, 2025).

Yet greater access does not, on its own, make journalism more reliable: the more material there is, the harder it becomes to find the one detail that matters (Yimam et al., 2016). When *The Atlanta Journal-Constitution* investigated sexual misconduct by doctors, a machine-learning program narrowed more than 100,000 disciplinary documents to around 6,000 likely cases, but reporters still read every case before drawing conclusions (Atlanta Journal-Constitution, 2016). When that check is skipped, the consequences show up in print: in 2025, the *Chicago Sun-Times* published a summer reading list in which ten of the fifteen recommended books did not exist, after a freelance writer used AI to produce it without verifying the results (Harrington, 2025). This reflects a limitation major AI developers acknowledge: models can present false information confidently and struggle to communicate uncertainty (Anthropic, n.d.; Google, n.d.; Kalai et al., 2025). Newsrooms therefore need to distinguish what AI suggests from what journalists have independently established.

If AI's efficiency becomes an excuse to cut reporting and verification, the risk only grows. Doctorow's (2023) concept of enshittification, describing how platforms deteriorate as their value is redirected toward commercial interests, offers an analogy: AI could increase

journalistic output while weakening the practices that make it valuable. Early signs are already visible: Graphite estimated that roughly half of sampled English-language articles published in early 2026 were primarily AI-generated (Paredes et al., 2026), while NewsGuard had identified 1,121 AI-generated news websites operating with minimal human editorial oversight as of November 11, 2024 (O'Connor, 2024).

These changes coincide with uncertainty about authorship and weakening confidence in news. In September 2026, NOS reported on AI assistance in social-media messages published by Dutch prime minister Rob Jetten and his party, prompting debate about the authenticity of apparently personal communication (NOS, 2026). For news organisations, questions about who produced a text also concern who checked its claims and accepts responsibility for publication. This is particularly pressing in the Netherlands, where trust in news reached its lowest recorded level in 2025 since measurements began in 2018 (Commissariaat voor de Media, 2025). Although this does not show that AI caused the decline, it underscores the need for careful decisions about AI use.

Problem Statement

The policy problem is how European newsrooms can expand their investigative capacity while preserving verification, professional judgement, and editorial accountability. Stray (2019) identifies data preparation and record linkage as promising applications of AI, but emphasises the extensive checking required for journalistic inference and the difficulty of translating judgements about newsworthiness into computational rules. Increasing the capacity to process information does not resolve the editorial task of determining what that information supports. This paper therefore examines what journalists need to evaluate a system's suitability, detect overlooked evidence, and test proposed interpretations, and what time, support, and authority let them act on their assessments. This grounds two complementary recommendations: appointing an AI and verification specialist and adopting a written newsroom AI policy.

Analysis

AI Literacy and Task-Specific Evaluation

For Deuze and Beckett (2022), AI literacy is more than knowing how to operate a tool: it includes understanding AI, recognising where it can contribute to journalism and when it should be avoided, and helping colleagues develop the skills to use it, along with ethical judgement about whose interests a system serves and how its use affects journalists and audiences. Stray (2019) shows why those judgements must be specific to the investigation. Research into Donald Trump's property dealings required reconstructing ownership and debt relationships across complex contracts and records, yet building an AI system for it was difficult because training examples were scarce, costly, and unlikely to generalise.

In practice, AI literacy should enable journalists to evaluate a tool against a particular investigation's demands: defining what it must do, testing representative material, and identifying errors that could undermine their findings. Its practical test is whether journalists can explain why a tool is suitable, recognise its limitations, and decide when further checking or another method is needed.

Editorial Judgement in Selection and Verification

Bellingcat's civilian-harm project applies this to less structured material, limiting the system to identifying posts for further examination. Ramalho and Waters (2026) describe an experimental XGBoost classifier that ranked Telegram posts by their likelihood of indicating civilian harm in Ukraine, using text and metadata rather than the images themselves. Researchers reviewed the ranked posts, preserving selected material through Auto Archiver and verifying it manually in ATLOS, so they, not the classifier, established what the evidence showed. This could reveal false positives but not false negatives: relevant posts the system overlooked.

To account for this, Bellingcat randomly flagged posts that no model had selected, alongside examining influential features and comparing historical patterns (Ramalho & Waters, 2026), consistent with Webber's (2013) approach of sampling both retrieved and unretrieved documents. Because missed evidence may leave incidents or communities overlooked, journalists need to check both what a tool selects and what it leaves out.

These checks raise a prior question: what should the system treat as relevant? BBC investigators examining a Russian military unit used a language model to rank tens of thousands of social-media posts against criteria including battlefield conditions and contradictions of official accounts, then concentrated on the highest-ranked 5–10% (Online News Association, 2024). This practitioner account does not quantify missed material, but it shows how editorial priorities became ranking instructions that shaped what investigators read. Journalists must therefore assess whether the criteria serve the investigation, not only whether the system applies them, since narrow criteria applied consistently can still exclude relevant evidence.

When AI proposes an interpretation rather than selecting material, evaluation must also test the basis of its conclusion. Postma and Patin's (2025) Bellingcat evaluation comprised 500 geolocation tests, with 20 models examining the same 25 previously unpublished photographs. Plausible explanations sometimes accompanied incorrect conclusions: one model interpreted a feature as Thai script and proposed Bangkok for a photograph taken in Singapore. The exercise cannot establish the reliability of later systems, but it shows that a reasoned explanation may rest on a mistaken observation. Journalists must test the observations and assumptions behind such an output against independent geographical evidence; the explanation itself cannot substitute for corroboration.

Organisational Conditions for Editorial Control

These checks require newsroom time and resources as well as individual expertise. Simon's (2024) research, drawing on interviews with 134 news workers across 35 organisations and 36 additional experts, finds that efficiency gains depend on task and context and that saved time may be absorbed by additional production demands. Faster processing need not create more time to test selection methods, investigate omissions, or corroborate interpretations. Journalists also need authority to act on what these checks reveal: identifying an unreliable selection or unsupported interpretation has limited value unless that assessment can delay, redirect, or halt reliance on the output.

Newsrooms also shape whether journalists acquire the necessary knowledge. In a 2024 survey of 1,004 UK journalists, 60% reported AI guidelines covering at least one

surveyed issue, while 32% reported employer-provided AI training; 50% of journalists at conglomerates reported such training, compared with 14% at independent outlets (Thurman et al., 2025). These figures neither describe Europe nor measure training effectiveness, but show that formal guidance can coexist with uneven access to practical preparation. Read alongside Deuze and Beckett (2022), they suggest responsible judgement cannot be solely an individual obligation; organisations must provide opportunities to develop and apply the required knowledge.

The analysis therefore identifies two interdependent requirements: AI literacy to evaluate systems, and organisational support (training, time, expertise, and authority) to act on those evaluations.

Policy Options

Table 1 compares five approaches on whether each provides both the expertise to assess AI-assisted work and the procedures to respond when problems arise. Its assessments are reasoned judgements, not measured rankings or cost estimates.

Table 1

Comparison of Newsroom Governance Options

Option	Investigative usefulness	Verification	Accountability	Resources
Unmanaged adoption	Flexible experimentation	Depends on individual practice	AI-specific decisions may lack records	Few added administrative requirements
Broad restriction	Forgoes prohibited uses	Avoids their errors; retains manual checking	Clear boundary if scope is defined	Requires monitoring and alternatives
Specialist alone	Supports task selection	Provides technical assessment	May depend on one person	Requires allocated expert time
Editor-led policy	Permits bounded uses	Limited by staff competence	Makes expectations explicit	Requires drafting and review
Editor-led policy and specialist	Supports supervised use	Links assessment with procedures	Records decisions and responsibilities	Requires coordination, time and money

Unmanaged adoption allows free experimentation with few administrative requirements, but leaves verification and record-keeping to individual practice. A reporter may check what a system selected without examining what it missed, and with decisions unrecorded, colleagues have no agreed way to question the method.

Restriction gives up prohibited uses and their errors while manual checking continues. Its boundary is only clear if its scope is defined, and it need not rule out all experimentation. The Associated Press's 2023 guidance, for example, allowed cautious experimentation with ChatGPT while advising staff against entering confidential information and requiring generated outputs to be treated as unvetted source material (Barrett, 2023). Restrictions also need monitoring and workable alternatives. Microsoft and LinkedIn (2024) found that 52% of workplace AI users were reluctant to disclose using it for their most important tasks. The finding spans sectors and does not show that restrictions cause concealment, but it cautions against assuming that a ban stops use. Editors should explain restrictions and provide approved tools or manual alternatives, keeping experimentation within what the newsroom can supervise.

A specialist working alone can help select suitable tasks and provide technical assessment. Nishal et al. (2024) propose assessing journalistic AI on output quality, fit with newsroom work processes, and alignment with professional values, so a general performance score cannot settle whether a tool suits a particular investigation. A specialist could design tests on these dimensions and investigate recurring errors, but the commissioning editor would still decide whether the evidence supports the investigation's claims. Without agreed procedures, however, oversight may depend on one person, and specialist advice may be followed inconsistently or mistaken for publication approval.

An editor-led policy permits bounded uses and makes expectations explicit by placing approval within existing editorial lines, which can work well when AI is used occasionally or for clearly defined tasks. The Council of Europe's Steering Committee on Media and Information Society (CDMSI, 2024, sections 1.2 and 1.6) frames AI adoption as an editorial decision and accepts that smaller organisations may give oversight to someone in an existing role. Under Chicago Public Media's (2026) policy, staff need supervisor approval before

using AI, must be able to explain its contribution, and can raise concerns with an AI oversight committee. Such a policy must specify who authorises a task, what testing is recorded, and when unresolved concerns require suspension. Its verification, however, is only as strong as staff competence: a testing requirement achieves little if nobody can carry the tests out. The CDMSI (2024, section 3.5) accordingly calls for ongoing training that brings journalists and technologists together.

Pairing an editor-led policy with specialist support covers the weaknesses of each. The specialist's findings feed a recorded decision, and when a weakness turns up, there is an agreed route to fixing it or stopping. Because the record holds both the specialist's assessment and the editor's approval, responsibility is documented and shared rather than resting on one person.

The arrangement is not free. Testing, consultation, documentation and training compete with time for reporting; specialist review could slow decisions, and reporters might grow less willing to challenge technical advice. These are risks to manage, not inevitable outcomes, so reviews should focus on consequential or unfamiliar uses, with cover arrangements and training that lets reporters handle routine checks.

Recommendation A: Appoint a Specialist

Editors-in-chief should designate a specialist who reports to a senior editor, with protected time and access to technical advice. The role involves assessing AI applications, training staff, and advising on any use that could shape editorial decisions. Some newsrooms will create a dedicated post; others will assign the duties to a technically skilled staff member or share a specialist with partners.

The specialist's work starts before adoption. With the commissioning editor, they should define the task, agree what the tool must do reliably, test it on already-verified material, and record its limitations. A transcription tool must capture names, numbers and quotations accurately, including in noisy recordings or unfamiliar accents. The decision (whether the tool is approved, for which tasks, and why) belongs in an assessment register and should be revisited whenever the provider updates the tool or changes how it handles uploaded information. During reporting, the specialist is consulted when an output is disputed

or checks must go beyond a tool's selected results, with a named second contact so decisions never stall.

Training should be practical and tied to approved applications. Staff should practise spotting unsupported inferences, checking whether an output is faithful to its source, and recognising when a question needs further review, with performance judged on these tasks rather than self-reported confidence. Because the CDMSI (2024, sections 3.1 and 3.5) warns that dedicated AI roles should not erode other staff's everyday skills, building colleagues' capability is part of the specialist's job.

None of this shifts responsibility. Reporters remain answerable for their evidence, and publishing editors for approval; technical advice is not editorial authorisation. If an application produces serious errors or exposes sensitive information, the specialist can recommend suspending it immediately, but the designated senior editor decides, after investigation, whether it can be used again. If colleagues come to lean too heavily on the specialist, clear records and good training give them the means to question decisions themselves.

Recommendation B: Adopt a Written AI Policy

The editor-in-chief should approve a written AI policy covering staff, freelancers and collaborative investigations, setting out which uses are permitted, what has to be documented, how concerns are escalated, and who is responsible for publication decisions. The controls below are grouped by journalistic function; in each, journalists must also judge the information's sensitivity and the harm its use could cause.

For transcription, translation and summarisation, journalists should check key details, quotations and meaning against the original and confirm, before uploading, that the tool is approved for that kind of information. Where AI is used to filter documents or rank posts, the selection criteria and their editorial purpose should be recorded, and journalists should check both the selected results and a sample of what was left out. Suggested locations, connections or explanations should be treated as leads, not findings, and tested against independent evidence, with a record of what supports or contradicts them. The strictest controls apply when AI proposes text, headlines or summaries for publication: factual claims must be

verified against original sources, attribution, context and expressed uncertainty checked, and the responsible editor must approve the material and decide whether to disclose its AI origin. These controls attach to the task, not to the tool. Before using a model to rank leaked records, for example, the reporter should document the method and obtain the commissioning editor's approval, with specialist advice where needed.

Recommendation C: Run a Twelve-Month Pilot

Months 1–2: setup and baseline. The accountable editor should select limited tasks, justify AI use, choose the oversight model, allocate time and train participants. Using previously checked material, measure existing errors and checking costs, have independent reviewers reconcile a subset of records into a reference set, and agree which errors would make results unusable.

Months 3–6: logging. Record processing, verification, correction and consultation time; consequential errors; missed records found through sampling; and substantiated leads. Assess competence through practical exercises, track all costs including displaced reporting time, and have an independent reviewer audit sampled cases.

Month 6: review. The editor-in-chief, reporters and specialist should compare results with the baseline, report sampling limits and failures by language or document type, and narrow approval where necessary. Where feasible, compare the same task under editor-led oversight to assess the specialist's contribution.

Months 7–11: refinement. Test agreed improvements without introducing unassessed uses.

Month 12: decision. The editor-in-chief should continue uses only where findings are defensible, benefits justify total workload, and checks and staffing are adequate. Stop uses where confidentiality failures persist or consequential errors cannot reliably be detected, adopt simpler oversight where it suffices at lower cost, and publish a brief account that protects confidential information. These criteria support a local continuation decision, not a general claim about European journalism.

Conclusion

A specialist is worth having when they solve recurring evaluation problems that editors cannot handle on their own. If a simpler arrangement gives the same scrutiny, a specialist wastes resources, and no specialist is enough if the newsroom cannot protect its sources or catch serious errors. A written policy turns these limits into enforceable rules. The pilot should show which uses remain worthwhile once the full cost of verification, checks for missed evidence and oversight is counted. Any expansion should rest on that evidence, not on the assumption that faster processing makes for better investigative journalism.

References

Anthropic. (n.d.). *Reduce hallucinations*. Claude Platform Docs. Retrieved October 16, 2026, from https://platform.claude.com/docs/en/test-and-evaluate/strengthen-guardrails/reduce-hallucinations

Atlanta Journal-Constitution. (2016). *Doctors & sex abuse: About the AJC's investigation of doctor misconduct*. https://doctors.ajc.com/about_this_investigation/

Barrett, A. (2023, August 16). *Standards around generative AI*. Associated Press. https://www.ap.org/the-definitive-source/behind-the-news/standards-around-generative-ai/

Center for News, Technology & Innovation. (2026, June 15). *AI applications in investigative journalism*. https://cnti.org/briefing/ai-applications-in-investigative-journalism/

Chicago Public Media. (2026, February 13). *Acceptable use of artificial intelligence (AI) policy in the Chicago Public Media newsroom*. Chicago Sun-Times. https://chicago.suntimes.com/ai-policy

Commissariaat voor de Media. (2025). *Digital News Report Nederland 2025*. https://www.cvdm.nl/documents/3284/Digital-News-Report-2025-1.pdf

Deuze, M., & Beckett, C. (2022). Imagination, algorithms and news: Developing AI literacy for journalism. *Digital Journalism*, *10*(10), 1913–1918. https://doi.org/10.1080/21670811.2022.2119152

Doctorow, C. (2023, January 23). The 'enshittification' of TikTok. *Wired*. https://www.wired.com/story/tiktok-platforms-cory-doctorow/

Doumouro, C. (2025, May 14). *How we use machine learning to find passports and unlock one key to offshore secrecy*. International Consortium of Investigative Journalists. https://www.icij.org/inside-icij/2025/05/how-we-use-machine-learning-to-find-passports-and-unlock-one-key-to-offshore-secrecy/

Google. (n.d.). *What are AI hallucinations?* Google Cloud. Retrieved October 16, 2026, from https://cloud.google.com/discover/what-are-ai-hallucinations

Harrington, A. (2025, May 20). “Summer reading list” with AI-generated titles of books that don’t exist runs in Chicago Sun-Times. *CBS News Chicago*. https://www.cbsnews.com/chicago/news/summer-reading-ai-generated-titles-chicago-sun-times/

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why language models hallucinate* (arXiv:2509.04664). arXiv. https://doi.org/10.48550/arXiv.2509.04664

Microsoft & LinkedIn. (2024, May 8). *AI at work is here. Now comes the hard part* (2024 Work Trend Index annual report). Microsoft. https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part

Nishal, S., Li, C., & Diakopoulos, N. (2024). *Domain-specific evaluation strategies for AI in journalism* (arXiv:2403.17911). arXiv. https://doi.org/10.48550/arXiv.2403.17911

NOS. (2026, September 10). *Premier Jetten liet zijn berichten door AI schrijven: Handig of zorgelijk?* [Prime Minister Jetten had his posts written by AI: Useful or worrying?]. https://nos.nl/artikel/2630471-premier-jetten-liet-zijn-berichten-door-ai-schrijven-handig-of-zorgelijk

O’Connor, H. (2024, November). *Watch out: AI “news” sites are on the rise*. NewsGuard. https://www.newsguardtech.com/insights/watch-out-ai-news-sites-are-on-the-rise/

Online News Association. (2024, October 22). *Case study: Using AI to analyze open-source intelligence in Ukraine war reporting*. https://www.journalists.org/news/case-study-using-ai-to-analyze-open-source-intelligence-in-ukraine-war-reporting

OpenAI. (2025, February 2). *Introducing deep research*. https://openai.com/index/introducing-deep-research/

Paredes, J. L., Druck, G., Benson, B., & Smith, E. (2026, May 15). *AI now writes as many online articles as humans*. Graphite. https://graphite.io/five-percent/research/ai-now-writes-as-many-online-articles-as-humans-do

Postma, F., & Patin, N. (2025, June 6). *Have LLMs finally mastered geolocation?* Bellingcat. https://www.bellingcat.com/resources/how-tos/2025/06/06/have-llms-finally-mastered-geolocation/

Ramalho, M., & Waters, N. (2026, June 25). *How to use AI to help find civilian harm*. Bellingcat. https://www.bellingcat.com/resources/2026/06/25/how-to-use-ai-to-help-find-civilian-harm-conflict-report-monitor-war-machine-learning-telegram/

Simon, F. M. (2024). *Artificial intelligence in the news: How AI retools, rationalizes, and reshapes journalism and the public arena*. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/ncm5-3v06

Steering Committee on Media and Information Society. (2024). *Guidelines on the responsible implementation of artificial intelligence (AI) systems in journalism* (CDMSI(2023)014). Council of Europe. https://rm.coe.int/cdmsi-2023-014-guidelines-on-the-responsible-implementation-of-artific/1680adb4c6

Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism*, *7*(8), 1076–1097. https://doi.org/10.1080/21670811.2019.1630289

Thurman, N., Thäsler-Kordonouri, S., & Fletcher, R. (2025). *AI adoption by UK journalists and their newsrooms: Surveying applications, approaches, and attitudes*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-ea11-q402

Webber, W. (2013). Approximate recall confidence intervals. *ACM Transactions on Information Systems*, *31*(1), Article 2. https://doi.org/10.1145/2414782.2414784

Yimam, S. M., Ulrich, H., von Landesberger, T., Rosenbach, M., Regneri, M., Panchenko, A., Lehmann, F., Fahrer, U., Biemann, C., & Ballweg, K. (2016). new/s/leak – Information extraction and visualization for investigative data journalists. In *Proceedings of ACL-2016 System Demonstrations* (pp. 163–168). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-4028