Harvard Kennedy School Misinformation Review[1]
August 2026, Volume 7, Issue 5
Creative Commons Attribution 4.0 International (CC BY 4.0)
Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-208
Website: https://misinforeview.hks.harvard.edu

Commentary

Turning OSINV inside the machine: Open-source investigations and information disorder

Open-source investigators and OSINV journalists have spent the last decade turning publicly available data into evidence of war crimes, environmental harms, and human rights abuses. More recently, they have trained the same lens on a different target: the algorithms, bot networks, and generative AI systems that industrialise deception and information disorder. Moving beyond debunking and verification, this reporting investigates the infrastructures that make falsehood profitable and visible, a turn four case studies here demonstrate.

Authors: Philip Di Salvo
Affiliations: Institute for Media and Communications Management (MCM), University of St. Gallen, Switzerland
How to cite: Di Salvo, P. (2026). Turning OSINV inside the machine: Open-source investigations and information disorder. *Harvard Kennedy School (HKS) Misinformation Review*, *7*(5).
Received: June 2nd, 2026. Accepted: July 31st, 2026. Published: August 12th, 2026.

Introduction

Open-source investigations (OSINV)—a term increasingly preferred by journalists and members of international civil society over "open-source intelligence" (OSINT) to avoid the military connotations of the latter (van der Woude et al., 2024)—are based on the use of publicly available digital material and data to investigate events, verify claims, and produce journalistic evidence. OSINV journalists have opened routes to evidence and accountability that conventional reporting methods cannot easily reach, particularly when distance, danger, or institutional opacity stand in the way (Dubberley et al., 2020). The clearest expressions of this shift have so far concerned the physical world: the geolocation of imagery, the verification of footage, the reconstruction of chains of events in conflict zones, and the documentation of human rights violations and environmental harm.

Yet, the analytical reach of OSINV is not exhausted by its engagement with offline events. This practice belongs to a broader family of digital investigations: an applied mode of inquiry concerned with the verification of online content, the exposure of disinformation and digital deception, and the tracing of the actors and infrastructures that sustain manipulated media (Silverman, 2020). Thus, OSINV's repertoire can extend beyond visual reconstruction, and a consequential new direction is now coming into view: OSINV's deployment against the infrastructures, technologies, and actors of information disorder themselves. OSINV is conducted not within single newsrooms alone but also across hybrid networks that link established journalistic institutions to investigative collectives and forensic organisations, each

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

contributing distinct competencies to a shared verification effort (Reese & Chen, 2022). Organisations such as Bellingcat thus operate adjacent to conventional journalism rather than within it, with their investigations frequently reaching publication through partnerships with established news institutions. OSINV practices are also deployed to verify evidence and source materials emerging from the contexts OSINV investigations report on, offering a crucial contribution to fact-checking, debunking, and event reconstruction, where propaganda and the "fog of war," the uncertainty, confusion, and lack of reliable information that people face during conflict or rapidly changing situations, allow dis- and misinformation to thrive.

As Dodds et al. (2025) argue, OSINV has been institutionalised within the journalistic field and now constitutes an established journalistic genre. It has drawn on satellite imagery to scrutinise locations from above and interrogated visual evidence on digital platforms to verify footage, geolocate images, and reconstruct chains of events, equipping the profession with an additional layer of "witnessing" capacity (Tait, 2011). OSINV has since been taken up across beats other than war and crisis reporting and integrated into major and mainstream newsrooms reporting with in-house teams, notably in corruption cases and financial journalism, where the investigation of online traces has become central to the reporting itself. This Commentary takes the expansion of OSINV as its point of departure, with a twofold aim: first, to position OSINV journalism as a distinctive response to the contemporary misinformation environment; and second, to examine the emerging empirical cases that have thus far received limited scholarly attention.

The digital traces of information disorder as misinformation evidence

In the contemporary information ecosystem, mis- and disinformation, propelled by algorithmic amplification (Özturan et al., 2025) and opaque platform architectures (Hu, 2020), erode democratic discourse and journalistic authority. This disorder is now compounded by generative AI (Park & Nan, 2025) and the proliferation of "AI slop," which deposits a further synthetic stratum of informational pollution in the form of "slopaganda" (Klincewicz et al., 2025). Yet, this expanding information disorder invariably leaves digital traces that journalists may mobilize as evidence (Hepp et al., 2018): the infrastructural fingerprints of websites and domains (Williams & Carley, 2023), the financial flows that sustain disinformation operations (Diaz Ruiz, 2025), the provenance signals embedded in synthetic media, the account-level patterns indicative of coordinated inauthentic behavior (Giglietto et al., 2020), and the artefacts of generative systems themselves. Read in the spirit of Schneier's (2015) observation that data is the by-product of computation, these traces constitute the empirical terrain upon which OSINV journalism can be brought to bear, transforming the architecture of online deception into something that can be observed, documented, and held to account.

This availability remains only partially exploited. OSINV has proven its value chiefly in debunking false claims and tracing manipulated media, while the systems that produce, amplify, and profit from deception have attracted little scrutiny. To reorient the field toward those systems is to effect a shift at once methodological, epistemological, and political: a passage from OSINV as fact-checking to OSINV as infrastructure investigation, intelligible because both draw on a shared culture of verification, yet decisive because it moves the unit of analysis from the discrete claim to the system that sustains falsehood. This reorientation finds a scaffold in Richardson's (2024) *Nonhuman Witnessing*, which holds the human witness inadequate to the scale and opacity of contemporary harm and proposes in its place a framework in which algorithmic entities themselves bear witness. Algorithmic opacity, on this view, is not incidental to the disinformation problematic but constitutive of it, foreclosing accountability in the service of commercial and strategic interests. To investigate such systems is therefore to elicit a form of nonhuman witnessing, compelling opaque infrastructures to yield their data exhaust as testimony.

A substantial practitioner literature supplies the methodological substrate for this work. Its canonical text is Silverman's (2020) edited *Verification Handbook: For Disinformation and Media Manipulation*, which gathers contributions from investigators at Bellingcat, ProPublica, and other leading outlets. It is complemented by the Global Investigative Journalism Network's (GIJN) (2023) *Reporter's Guide to Investigating Digital Threats*—which holds, in operational terms, that any actor operating online leaves recoverable traces—and by Bellingcat's continuously updated Online Investigation Toolkit, which catalogues the instruments through which such traces are extracted.

OSINV inside the machine: A taxonomy

How does this reorientation of OSINV toward scrutinising the computational infrastructures and automated processes that underpin contemporary information disorder take shape in practice? What follows is a purposive selection of international case studies, assembled to illustrate rather than to sample. Each case was chosen because it is well documented, widely reported, and exposes a distinct stratum of the disinformation apparatus, so that together they trace the phenomenon's contours rather than measure its extent. They are best arranged as a taxonomy keyed to the stratum of the disinformation apparatus each renders legible: the *generative*, where content is manufactured; the *distributive*, where it is circulated; the *economic*, where it is monetized; and the *agentive*, where the apparatus is re-attached to the agents who operate it. These strata are analytically separable but not mutually exclusive; in practice a single operation may manufacture, circulate, and monetise deception at once. The cases are therefore offered as demonstrations of what open-source investigation can render legible at each layer, rather than as an exhaustive partition of the field.

At the *generative* layer stands the investigation of AI-generated propaganda on TikTok, the "slopaganda" named earlier. The non-profit AIForensics, with the assistance of a network of media partners (including *The Guardian* and *Der Spiegel*), combined automated monitoring with metadata analysis to surface large-scale synthetic campaigns, examining several hundred autonomous accounts across some twenty languages, whose tens of thousands of AI-produced posts drew billions of views, of which only a fraction were labelled as synthetic.[2] By reading visual artefacts, posting patterns, and engagement metrics, investigators separated automated propaganda from authentic expression, disclosing both the velocity of fabrication at its origin and the insufficiency of claim-by-claim checking against content produced at industrial volume.

At the *distributive* layer, the *Doppelgänger* investigation reconstructs an apparatus of circulation rather than adjudicating its outputs. [3] The campaign, a pro-Kremlin operation active across Europe since at least 2022, propagated falsehood through clones of legitimate mastheads, sustained by look-alike domains and inauthentic accounts. Investigators mapped the enabling infrastructure: the German investigative newsroom CORRECTIV, with the forensics organization Qurium, traced the redirection chain a browser follows to a cloned site through a Ukrainian service provider to footprints leading toward the Russian Ministry of Defence,[4] findings the DFRLab and the U.S. Department of Justice later corroborated, the latter seizing some thirty domains. The inquiry exposed a systematic architecture of coordination and cross-platform synchronization.

At the *economic* layer, ProPublica's investigation of Google's advertising network shows disinformation sustained as a political economy.[5] Reverse-engineering an opaque system, reporters

[2] See: https://aiforensics.org/work/gen-ai-slop
[3] See: https://www.disinfo.eu/doppelganger-hub/
[4] See: https://correctiv.org/en/fact-checking-en/2024/07/22/inside-doppelganger-how-russia-uses-eu-companies-for-its-propaganda/
[5] See: https://www.propublica.org/article/google-alphabet-ads-fund-disinformation-covid-elections

reconstructed the company's concealed roster of publishers, matching most of its seller list to specific domains and apps. They found revenue-optimized algorithms placing major brands' advertisements on piracy operations, fraudulent traffic schemes, and disinformation sites, in several cases despite the company's own data indicating copyright infringement and persisting until reporters made contact. Its purchase here lies in showing that automated advertising rewards engagement irrespective of veracity, subverting the controls the platform professes to enforce.

At the *agentive* layer, the Mr. Deepfakes investigation closes the sequence by re-anchoring an opaque apparatus to an identifiable operator.[6] Confronting what had been the foremost marketplace for non-consensual synthetic pornography, a site that accrued billions of views before its 2025 closure, Bellingcat, with the Canadian Broadcasting Corporation and the Danish outlets *Politiken* and Tjekdet, penetrated the anonymity of its administration. That the investigation was carried out by an investigative collective operating adjacent to journalism, such as Bellingcat, in concert with a public-service broadcaster and two national newspapers is itself illustrative: it exemplifies the hybrid mode through which OSINV is increasingly conducted, in which classic journalistic institutions and investigative organisations combine their respective capacities to produce and publish verified evidence. Cross-referencing publicly available credential-breach data, investigators assembled from burner emails, recurring usernames, a distinctive password, and IP addresses a decade-long trail, corroborated by shared analytics tags and back-end software across associated domains. In surfacing both the agent and the enabling systems, the case shows how generative technologies are weaponized for gendered harassment, and how method can establish accountability where perpetrators shelter behind technical complexity and jurisdictional ambiguity.

The four investigations constitute but a stratigraphy of the disinformation apparatus—*manufacture*, *circulation*, *monetization*, *agency*—each layer rendered amenable to evidence and, ultimately, to account.

Conclusion

OSINV first proved itself as an instrument of accountability from above, trained on terrain, war zones, and sites of atrocity, environments of institutional opacity that cannot be reached in the first person. This Commentary has argued that the black-box character of contemporary algorithmic systems offers an analogous opening. Like the inaccessible terrain of a war zone, the proprietary recommendation engine, the coordinated bot network, and the generative model are opaque by design; yet they too leave traces across the open web, and those traces can be mobilized, through OSINV, as evidence with which to investigate them. By this means journalists can begin to address what happens inside the machine, exposing the otherwise obscure dynamics of datafication and algorithmic mediation. The four case studies examined here demonstrate the value of this turn, particularly in the investigation of information disorder. They intervene at the level of the apparatus itself: the dynamics that sustain information-disorder activities, the actors who orchestrate them, and the tools and infrastructures on which they depend. In rendering that apparatus legible, OSINV makes the machine testify against itself: its data exhaust becomes evidence, and the investigation itself a form of nonhuman witnessing.

To go inside the machine is to assign journalism a role of scrutinizing the infrastructures through which public communication operates, making the conditions of mediation themselves an object of reporting, along with the actors and practices that can disrupt these infrastructures through the use and misuse of algorithms and artificial intelligence. Where the digital ecosystem increasingly determines, through opaque computational means, which information reaches publics and on what terms, a journalism capable of rendering those mechanisms legible does more than expose falsehoods. It is in this sense that OSINV, directed at the architectures of deception rather than merely their outputs, emerges as a

[6] See: https://www.bellingcat.com/news/2025/05/07/canadian-pharmacist-linked-to-worlds-most-notorious-deepfake-porn-site/

distinctive and increasingly necessary response to the contemporary misinformation environment, and as a fitting subject for the sustained scholarly attention this Commentary has sought to invite.

Bibliography

Diaz Ruiz, C. A. (2025). Disinformation and fake news as externalities of digital advertising: A close reading of sociotechnical imaginaries in programmatic advertising. *Journal of Marketing Management, 41*(9–10), 807–829. https://doi.org/10.1080/0267257X.2024.2421860

Dodds, T., van der Velden, L., Torres, G., El-Masri, A., Reese, S. D., Fiorella, G., Farr, R. A. R., Ivens, G., & Kotišová, J. (2025). On the institutionalization of OSINV in journalistic practice. *Journalism & Mass Communication Quarterly, 102*(3), 617–641. https://doi.org/10.1177/10776990251334382

Dubberley, S., Koenig, A., & Murray, D. (Eds.). (2020). *Digital witness: Using open source information for human rights investigation, documentation, and accountability.* Oxford University Press. https://doi.org/10.1093/law/9780198836063.001.0001

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020). It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society, 23*(6), 867–891. https://doi.org/10.1080/1369118X.2020.1739732

Global Investigative Journalism Network. (2023). *GIJN reporter's guide to investigating digital threats.* https://gijn.org/resource/gijn-reporters-guide-to-investigating-digital-threats/.

Hepp, A., Breiter, A., & Friemel, T. N. (2018). Digital traces in context | Digital traces in context—An introduction. *International Journal of Communication, 12*, 439–449. https://ijoc.org/index.php/ijoc/article/view/8650

Hu, M. (2020). Cambridge Analytica's black box. *Big Data & Society, 7*(2), Article 2053951720938091. https://doi.org/10.1177/2053951720938091.

Klincewicz, M., Alfano, M., & Fard, A. E. (2025). Slopaganda: The interaction between propaganda and generative AI. *Filosofiska Notiser, 12*(1), 135–162. https://www.filosofiskanotiser.com/KlincewiczAlfanoFard.pdf

Özturan, B., Quintana-Mathé, A., Grinberg, N., Ognyanova, K., & Lazer, D. (2025). Declining information quality under new platform governance. *Harvard Kennedy School (HKS) Misinformation Review, 6*(3). https://doi.org/10.37016/mr-2020-176

Park, S., & Nan, X. (2025). Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *AI & Society, 41*(2), 1501–1515. https://doi.org/10.1007/s00146-025-02620-3

Reese, S. D., & Chen, B. (2022). Emerging hybrid networks of verification, accountability, and institutional resilience: The US Capitol Riot and the work of open-source investigation. *Journal of Communication, 72*(6), 633–646. https://doi.org/10.1093/joc/jqac030

Richardson, M. (2024). *Nonhuman witnessing: War, data, and ecology after the end of the world.* Duke University Press. https://doi.org/10.1215/9781478027782

Schneier, B. (2015). *Data and Goliath: The hidden battles to collect your data and control your world.* W. W. Norton & Company.

Silverman, C. (Ed.). (2020). *Verification handbook for disinformation and media manipulation* (3rd ed.). European Journalism Centre. https://datajournalism.com/read/handbook/verification-3

Tait, S. (2011). Bearing witness, journalism and moral responsibility. *Media, Culture & Society, 33*(8), 1220–1235. https://doi.org/10.1177/0163443711422460

van der Woude, M., Dodds, T., & Torres, G. (2025). The ethics of open source investigations: Navigating privacy challenges in a gray zone information landscape. *Journalism, 26*(10), 2184–2202. https://doi.org/10.1177/14648849241274104

Williams, E. M., & Carley, K. M. (2023). Search engine manipulation to spread pro-Kremlin propaganda. *Harvard Kennedy School (HKS) Misinformation Review, 4*(1). https://doi.org/10.37016/mr-2020-112

Acknowledgements
This Commentary originates from an online talk hosted by the University of Wisconsin–Madison's Public Tech Media Lab in late 2025. The author is grateful to Tomás Dodds for the invitation and to the participants for the discussion that inspired the piece.

Funding
No funding has been received to conduct this research.

Competing interests
The author declares no competing interests.

Copyright
This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.