

Before the Ballot

Data Poisoning Risks to Election

Shay Hershkovitz¹ | October 13, 2026

¹ Dr. Shay Hershkovitz is a national security researcher and practitioner specializing in emerging and dual-use technologies, with a particular focus on artificial intelligence, defense innovation, and intelligence transformation. His work examines how AI, data-driven systems, and novel technological capabilities reshape intelligence, military operations, and strategic decision-making. He is an adjunct professor in the MA program in Applied Intelligence at Georgetown University and a senior researcher (adjunct) at RAND.

Table of Contents

Executive Summary	4
1. Data poisoning: mechanisms and typology	7
1.1 What data poisoning is	7
1.2 Three ways in.....	8
1.3 Why elections are the exposed case.....	11
1.4 What the evidence shows, and what would change it	12
2. Methodology	15
2.1 Case selection.....	15
2.2 Recurring questions for case comparison	16
2.3 Reading the cases	18
2.4 Note on sources	18
3. The Pravda network: retrieval-tier poisoning at scale	19
3.1 Actor and attribution confidence	19
3.2 Scale, automation, and the AI targeting logic.....	20
3.3 Targeting, languages, and election cycles	21
3.4 Mechanism: retrieval-tier poisoning	22
3.5 Distribution and amplification	23
3.6 Documented impact.....	23
3.7 Detection, attribution, and defensive response	26
4. CopyCop / Storm-1516: a state-attributed network that feeds the same corpus	27
4.1 Actor and attribution confidence	27
4.2 Timeline and operational scale	28
4.3 Targeting and operator intent toward AI.....	28
4.4 Where CopyCop becomes a retrieval-tier vector	29
4.5 Distribution and amplification	29
4.6 Documented impact.....	29
4.7 Detection, attribution, and defensive response	30
5. Three national cases: Germany, Australia, Moldova 2025	32
5.1 Germany, February 2025 federal election.....	32
5.2 Australia, May 2025 federal election	32
5.3 Moldova, September 2025 parliamentary election	33
5.4 Cross-case findings.....	34
6. Lessons for Israel 2026	36
6.1 Israel is inside the target set, and no one has measured it.....	36

6.2 The exposure predates the campaign and outlasts it.....	36
6.3 Attribution will be contested	37
6.4 No Israeli body owns the retrieval tier	37
7. Recommendations	38
References	41

Executive Summary

Israeli citizens are scheduled to go to the polls on October 27, 2026, and mid-term elections will take place in the United States a week later. This report addresses a specific type of threat to artificial intelligence systems – data poisoning – and its potential implications for the election cycle.

Data poisoning is the deliberate introduction of false or manipulative information into the environment from which artificial intelligence systems learn or retrieve answers. Poisoning can occur during a model's training phase, in datasets used to adapt a model for specific tasks, or at the retrieval tier—namely, the sources a model consults in real time to answer a user's question. Unlike conventional cyberattacks, the attacker does not need to breach a system: publishing content on the open web and making it available for indexing and retrieval is sufficient.²

The primary case of retrieval-tier poisoning examined in this report is the Russian-linked Pravda network. Throughout 2024 and 2025, the network operated across more than 300 domains in dozens of languages, publishing millions of articles a year. Many of its sites attract virtually no human audience, and their structure is better suited for indexing, retrieval, and citation by artificial intelligence systems than for ordinary news consumption. For example, ahead of Australia's May 2025 federal election, a dedicated site was established for the Australian target audience. An audit conducted by NewsGuard for the Australian Broadcasting Corporation found ten leading models repeating false election claims in 16.6 percent of tested responses, with some directly citing the site.

There is also public evidence of Russian intent to influence AI systems themselves as part of influence operations. John Mark Dougan, an American deserter affiliated with Russian influence operations (and currently under European Union sanctions), stated in Moscow in January 2025 that disseminating Russian narratives through his network of domains could effectively change artificial intelligence worldwide.

It should be emphasized, however, that no empirical evidence has yet established that data poisoning has actually changed how voters vote. This should not be interpreted as proof of a lack of impact: even in conventional influence operations, establishing a causal link between content exposure and voter behavior is difficult. The more practical question is whether the artificial intelligence systems used by Israeli citizens are already exposed to data-poisoning influence operations and, if so, to what extent. To date, no systematic, longitudinal audit has examined the answers leading AI models provide to questions concerning Israeli elections in

² For further details, see: Shay Hershkovitz, *Introduction to Data Poisoning: Fundamentals, Threat Models, and National Security Risks*, Memorandum 259, Institute for National Security Studies (INSS), 2026. <https://tinyurl.com/5n9xbpsm>

Hebrew and Arabic. Such an audit is a fundamental prerequisite for evaluating the extent of exposure over time.

Nevertheless, it is reasonable to assume that such operations are already underway in Israel ahead of the elections. Israel has already been targeted by Russian and Iranian influence operations, and Russian activity has demonstrated the capability to operate in Hebrew and target Israeli audiences, including by impersonating local media outlets. Networks of fake profiles linked to Iranian influence activity have also been documented, building credibility over time and changing their messaging at politically sensitive moments. During the conflict between Israel and Iran in June 2025, extensive use of AI-generated content was also documented. All this holds true even though there is currently no public forensic evidence showing that Iranian actors have targeted AI training data or retrieval systems in the context of Israeli elections.

Addressing the threat of data poisoning requires advance preparation prior to elections, as well as sustained efforts afterward. Retrieval-tier infrastructure develops and accumulates over time, and content that has already been indexed does not necessarily disappear even if the original website has been removed. Therefore, measures are needed both to reduce exposure to manipulative information and to build a systematic capacity for detecting it.

First, a mechanism should be developed for monitoring and periodically auditing leading models in Hebrew and Arabic, to test whether they repeat false narratives known in the Israeli context, and whether they cite domains affiliated with hostile influence networks. These audits will help assess the extent of exposure to data poisoning, identify changes over time, and establish a situational picture to assist in formulating a response.

Second, the presence of official and reliable information in the retrieval environment should be strengthened by publishing clear, structured, and machine-readable information on official and verified domains. Artificial intelligence systems should be treated as consumers of information alongside the human public, while ensuring that reliable public information on democratic processes and sensitive public issues is made accessible to them.

In parallel, public awareness regarding the limitations of using artificial intelligence systems should be increased. Their answers to political questions are not necessarily a neutral synthesis of reliable sources, and they may also draw on information deliberately injected into the retrieval environment. The purpose of public outreach is not to undermine trust in these systems, but rather to encourage critical use, source verification, and cross-checking of information.

In the long term, an institutional, legal, and technological infrastructure is needed to address this phenomenon. Among other things, institutional responsibility for monitoring data poisoning and coordinating the response must be defined, the suitability of the legal framework for dealing with deliberate manipulation of open information sources must be examined, and transparency requirements for model developers regarding the sources underlying their answers must be considered. These measures are intended to establish a sustained capability

to detect data poisoning and mitigate its effects, enabling the threat to be addressed both routinely and during periods of heightened public and political sensitivity.

1. Data poisoning: mechanisms and typology

Israel votes on October 27, 2026. The Knesset dissolved on July 17, and in the months since, hostile content could be published, indexed, and left to accumulate in the information environment that language models draw on. It should be assumed that much of the content on which AI systems' answers to Israeli voters' questions may rely has already been published and indexed.

This report makes three arguments.

First, hostile content can reach an AI system at three different points: the large body of text a model is trained on, the smaller datasets used to shape how it answers, and the live sources it consults when a user asks a question. Each requires a different kind of access, and each calls for a different response.

Second, the evidence across those three points is uneven. Researchers have demonstrated all three attacks in the laboratory. This review identified documented use against a real election at one of them only: the manipulation of sources models consult at the moment of use. None has been shown to change how people voted. This report argues that the pattern is what should be expected from an attack designed to go unnoticed, and it sets out what evidence would show that argument to be wrong.

Third, no one has measured Israel's exposure. There is no audit of how leading AI models answer Hebrew or Arabic questions about Israeli politics. That gap is the report's principal finding about Israel, and the recommendations are built to close it.

1.1 What data poisoning is

Data poisoning means feeding an AI system corrupted information so that it produces the answers an adversary wants. The corruption can be planted in the material the system learns from, or in the sources it looks up when someone asks it a question.

Two features set this apart from the cyber-attacks security institutions are built to handle. The attacker breaks into nothing. The material is published in the open, and the target collects it through its own normal, entirely legitimate processes. And the system carries on working as usual, because a well-built attack changes behavior only on the narrow set of questions the attacker cares about. Everything else looks fine.³

³Hershkovitz, S. (2026). *Data poisoning primer: Foundations, threat models, and national security risks* (Memorandum 256). Institute for National Security Studies, §2.0–2.2.

There are two moments when this can happen. The first is during training, when a model learns from a large body of text; distortions planted there become part of the model itself and can survive later rounds of correction. In one experiment, attempts to train the problem out taught the model to hide it better instead.⁴ The second is at the moment of use. Most major AI systems now search the web while answering a question about current events. Content planted in those sources can shape the answer even though it was never part of what the model learned.

Attacks of this kind are built to go unnoticed, and this is a property of the method rather than a gap in current tools. Standard quality tests do not catch a poisoned model, because those tests never ask the question that sets off the hidden behavior. The planted material itself is written to read as ordinary text, so automated screening finds nothing unusual in it. A field of research called interpretability, which tries to work out what a model has actually absorbed and why it answers as it does, may eventually help. It offers nothing to election officials, platforms, or outside auditors today.

1.2 Three ways in

The three points where corrupted content can enter differ in what an attacker has to reach, how much material the attack takes, and who is in a position to stop it. The report refers to these as tiers, and uses the distinction throughout.

The training material. Foundation models are trained on enormous collections of text scraped from the open web. An attacker does not need any relationship with the company building the model. He needs only to place material where that company will later collect it. Researchers have shown that the owner of a website can guarantee specific pages will be picked up by Common Crawl, the web archive underlying most major models, for a few thousand dollars per page.⁵ The same study documented two methods that require no technical access at all. One times a malicious edit to a heavily copied source such as Wikipedia so that it lands just before a periodic snapshot is taken: the corrupted version is the one collected, while anyone visiting the page afterwards sees it looking normal. The other buys up expired web addresses that still appear in the reference lists of published datasets and replaces what sits at them, so that the version indexed and the version downloaded are no longer the same.

How much material this takes is the surprising part. A 2025 study by Anthropic, the UK AI Security Institute, and the Alan Turing Institute found that around 250 documents were enough to plant hidden behavior in models, and that the number stayed roughly the same as models grew larger.⁶ Two hundred and fifty documents is far below anything a practical review of

⁴Hubinger, E., Denison, C., Mu, J., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv:2401.05566*.

⁵Carlini, N., Jagielski, M., Choquette-Choo, C. A., et al. (2023). Poisoning web-scale training datasets is practical. *arXiv:2302.10149*.

⁶Souly, A., et al. (2025). Poisoning attacks on LLMs require near-constant number of poison samples. *arXiv:2510.07192*. The models tested ranged from 600 million to 13 billion parameters. The finding applies to a specific backdoor class under defined experimental conditions and should not be read as a universal threshold for narrative or election-related poisoning.

training material could pick out. The collections involved run to trillions of words. Defending this tier is work only model developers can do: knowing where their data came from, filtering sources, and excluding domains known to be hostile.

The tuning data. After the initial training, models are refined on much smaller and more carefully assembled datasets that shape how they respond to questions. Reaching these requires more: access to the tuning process itself, through a contractor, a public dataset, or a curated collection. What it buys is greater precision. In one study, 52 examples, roughly a tenth of one percent of the dataset, moved a model's rate of negative responses about a named politician from zero to about 40 percent.⁷ The shift applied to every question about that person, not only to some hidden signal, and the poisoned examples themselves were unremarkable: the model's performance on everything else was almost identical to a clean one. The same weakness extends to the human feedback used to align models, where corrupted judgements can change behavior without any dataset being altered.

The live sources. Most major AI systems now search the web while answering a question about current events, and many business systems consult internal document stores in the same way. This is the easiest of the three to reach, because in the open-web case the attacker only has to publish. It also takes the least material. Planting as few as five documents in a searchable collection produced the attacker's intended answer more than 90 percent of the time, across several different models.⁸ Responsibility here is split. Developers control how sources are ranked and filtered. But whether authoritative material exists at all in a given language, on a given topic, is a matter for states, public bodies, and civil society. This is the tier where the case examined in section 3 sits, and the report refers to it as retrieval-tier poisoning.

Taken together, the three tiers point to the same conclusion: an actor with modest resources can create real vulnerabilities in a model without controlling any significant share of the material it learns from, and a successful attack will not announce itself through poor performance.

This is also why the range of actors is wide. States and state-aligned groups have the patience and reach for sustained campaigns, but the low entry cost puts some form of this within range of criminal groups, corporate rivals, insiders with access to labelling or tuning pipelines, and individuals acting alone.⁹

Two further attack surfaces sit alongside these three, and neither applies to Israel today. Voting advice applications — the tools some countries use to match voters to parties — can be manipulated through their configuration and data inputs; testing against Smartvote, which

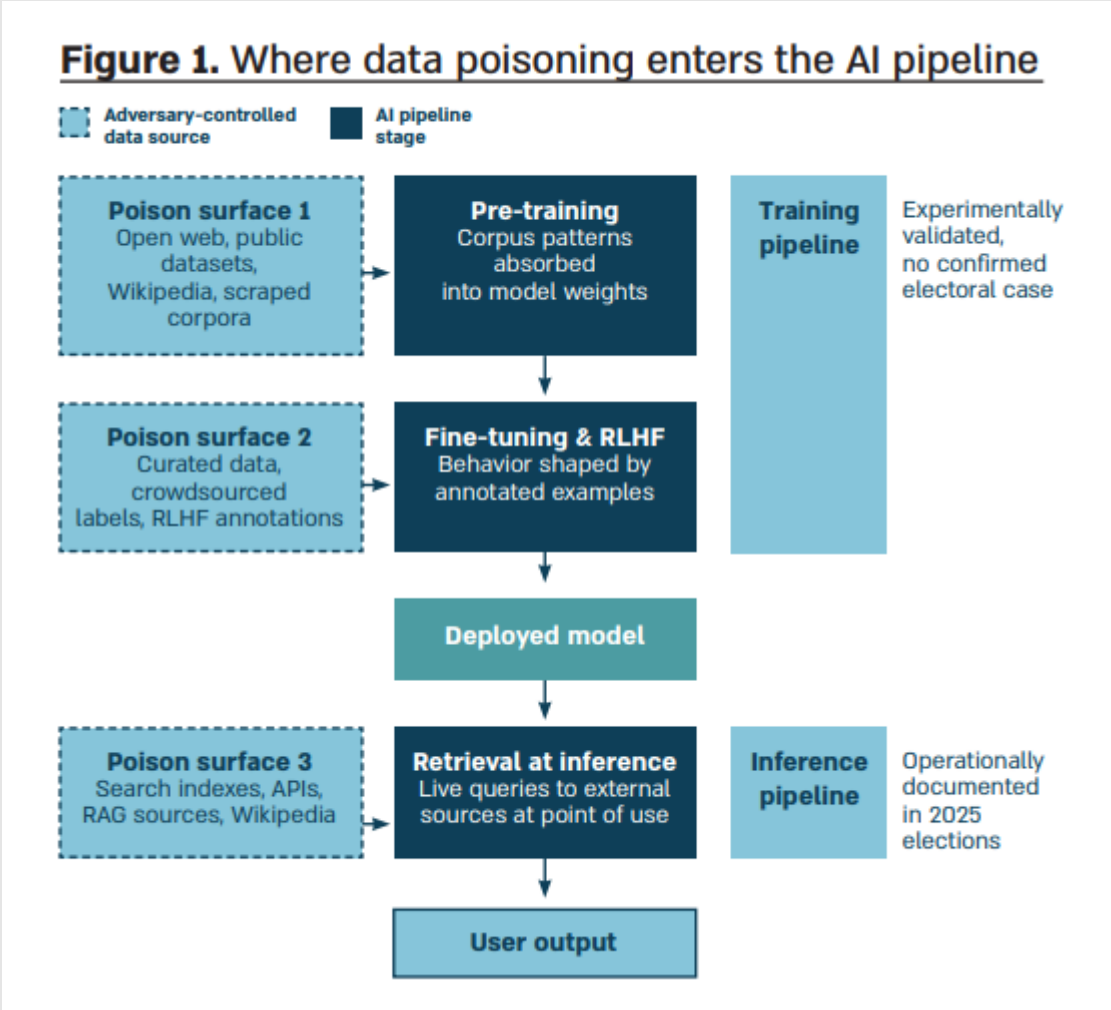
⁷Yan, J., Yadav, V., Li, S., et al. (2024). Backdooring instruction-tuned large language models with virtual prompt injection. *Proceedings of NAACL 2024*. <https://doi.org/10.18653/v1/2024.naacl-long.337>

⁸Zou, W., Geng, R., Wang, B., & Jia, J. (2024). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv:2402.07867*.

⁹HersHKovitz (2026), §2.6.

serves Swiss national elections, shifted the frequency with which particular parties were recommended by up to 261 percent.¹⁰ Election administration systems have shown related weaknesses: in one 2017 demonstration, fewer than 50 emails were enough to train a spam filter into blocking voting credentials.¹¹ Israel uses neither AI nor algorithmic recommendation in electoral administration, so neither surface is a live concern for this cycle, but both belong on the map.

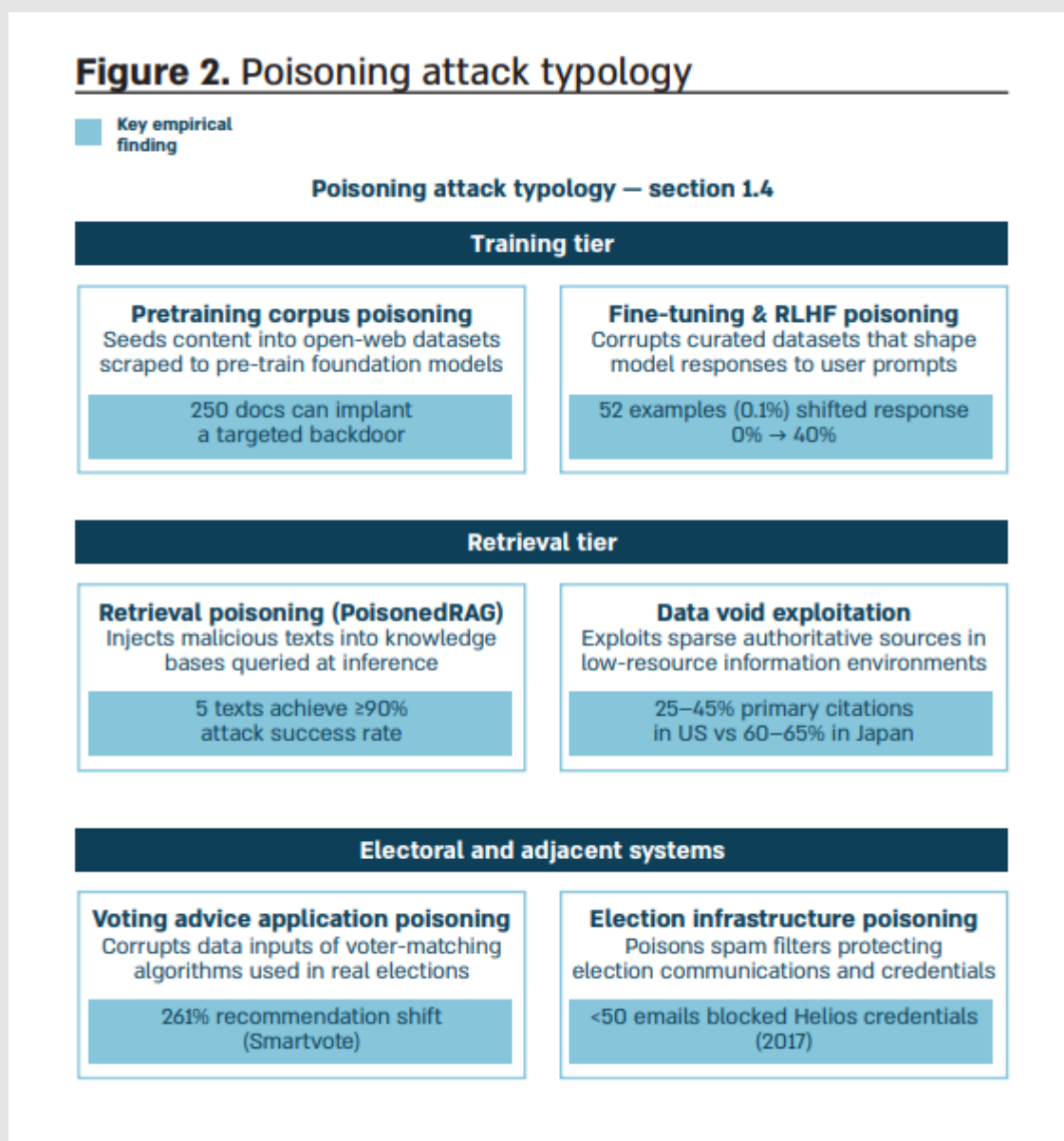
Figure 1. Where data poisoning enters the AI pipeline



¹⁰Berdoz, F., Brunner, D., Vonlanthen, Y., & Wattenhofer, R. (2025). Recommender systems for democracy: Toward adversarial robustness in voting advice applications. *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25)*. <https://doi.org/10.48550/arXiv.2505.13329> The 261 percent figure applies specifically to selective item inclusion, not to all three attack methods tested.

¹¹Jonker, H., Mauw, S., & Schmitz, T. (2017). Reverse Bayesian poisoning: How to use spam filters to manipulate online elections. *International Joint Conference on Electronic Voting*. https://doi.org/10.1007/978-3-319-68687-5_11

Figure 2. Poisoning attack typology



Two limits on scope. This report deals with the corruption of data and does not attempt to survey the wider field of attacks on machine learning. It also sets aside a related problem: tampering with a model directly, by altering its stored weights or interfering with software updates rather than by corrupting what it reads.

1.3 Why elections are the exposed case

Elections generate predictable, time-bounded demand for information. Voters, journalists, and AI systems all ask about the same things — candidate records, voting procedures, party platforms, counting timelines — within a compressed window. An adversary who places fabricated or distorted content in the sources those questions return does not need access to any model developer. It needs only to be present in the information environment before the questions arrive.

The material voters and machines draw on is also the same material. Candidate information, electoral procedures, party positions, and political history are documented primarily in news archives, Wikipedia, and publicly accessible reference sites — the exact sources that training collections scrape and that live retrieval queries at the moment of use.

What determines exposure is not the size of a language but how well covered a topic is by reliable sources. A 2025 study tested three leading commercial models on political questions and found that only 25 to 45 percent of the citations in answers about US politics came from primary sources, against 60 to 65 percent for Japan, with low-barrier sources supplying roughly 30 percent of citations. The authors attribute the gap to the density of the information ecosystem: the American political information space contains more question-specific gaps, leaving automated systems structurally more dependent on poor sources.¹² This is a measurement of where models draw their answers from, not a finding about poisoning. But it identifies the condition poisoning exploits — where reliable coverage is thin, whatever fills the gap is what gets retrieved.

For Israel the finding cuts both ways. Hebrew and Arabic are small-language environments, but that alone does not determine exposure. What matters is whether authoritative, machine-readable coverage of Israeli electoral topics — procedures, candidates, parties, contested political events — is dense enough to crowd out poor sources at the moment of retrieval. Israel's political information environment combines conditions that tend to produce gaps: contested and fast-moving topics, demand split across Hebrew, Arabic, Russian, and English, and official electoral information that is not published with machine retrieval in mind. Whether those gaps exist in practice is an empirical question no audit has yet answered, and closing it is the purpose of recommendation 7.

1.4 What the evidence shows, and what would change it

The evidence is uneven across the three tiers, and this section states where it stands before the case studies begin.

At the retrieval tier, operational use in an election context is documented. The Pravda network examined in section 3 was active around the 2024 European Parliament elections, the 2024 Georgian parliamentary elections, the 2025 German federal election, the 2025 Australian federal election, and the 2025 Moldovan parliamentary election, and independent audits have measured its content appearing in model outputs. At the training and tuning tiers, laboratory research has established that these attacks work, but this review identified no publicly documented use of either against an election. No case at any tier has been shown to change how people voted.

That gap — between what has been demonstrated in experiments and what has been caught in the wild — is the starting point for the report's argument.

¹²Mochizuki, R., Komatsu, S., Noguchi, S., & Ataka, K. (2025). Exposing citation vulnerabilities in generative engines. *arXiv:2510.06823*.

The absence of a confirmed case at the training and tuning tiers should not be read as evidence that nothing is happening there. Three properties of this attack class make documented electoral effects unlikely to surface even if poisoning is under way. The first, set out in section 1.1, is that the attack is built to evade routine detection: poisoned models pass standard tests, planted material reads as ordinary text, and hidden behavior can survive later attempts to correct it. The second is that even visible, large-scale election influence operations produce effects on individual voter behavior that are extremely hard to measure. Eady and colleagues found no measurable relationship between exposure to the Russian Internet Research Agency's 2016 US-election Twitter operation and attitudes, polarization, or voting behavior, despite three years of activity, billions of impressions, and a federal indictment.¹³ If operations conducted in the open consistently fail that evidentiary bar, demanding the same of an attack designed to be invisible sets a standard the influence-operations literature has already shown is rarely met. The third is that election authorities, platforms, and civil-society monitors lack the tools and the access to training data that detecting poisoning after deployment would require.

An argument that no possible evidence could disprove is worth little, so it is worth stating what would count against this one. If an audit of leading models on Israeli election questions in Hebrew and Arabic found no repetition of known hostile narratives and no citation of known hostile domains, that would indicate Israeli retrieval-tier exposure is low, and the case for treating it as a planning concern in this cycle would weaken accordingly. If repeated audits over the cycle showed exposure falling rather than rising, the same would follow. Recommendation 7 is designed to run that test. It has not been run, which is why this report describes an exposure that is plausible and unmeasured rather than one that is established.

Existing threat intelligence reflects a structural blind spot that this report is intended to address. The 2024–2025 assessments from Microsoft, Google, OpenAI, and Meta characterize AI-enabled election interference as producing low audience engagement and minimal observable persuasive effect; the US Director of National Intelligence in September 2024 described AI as "a malign influence accelerant, not a revolutionary influence tool."¹⁴ These assessments are sourced and credible, but they measure something else. Each concerns AI as a content-generation tool: synthetic media, AI-written articles, AI-cloned audio, all aimed at human readers and viewers. None addresses AI as a target. The stealth properties that make a successful poisoning operation hard to detect are precisely what would keep it out of threat reports built around audience-engagement metrics. A threat that does not announce itself in reach statistics will not appear in reports that measure reach.

¹³Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14, 62. <https://doi.org/10.1038/s41467-022-35576-9>

¹⁴Office of the Director of National Intelligence. (2024, September 23). Election security update [Press briefing transcript, 45-day pre-election update]. U.S. Department of State Foreign Press Center. <https://2021-2025.state.gov/briefings-foreign-press-centers/protecting-the-2024-election-from-foreign-malign-influence/>

The asymmetry between tiers also changes the time horizon for defense. Some threats appear during the campaign itself: synthetic video, fake accounts, paid amplification, spoofed election notices, coordinated messaging in encrypted groups. Others are already in place, in the sources AI systems retrieve or learn from. Polling-day security, platform takedowns, and post-hoc fact-checking remain necessary, but they do not address contamination that entered the reference environment months or years before an election. Content seeded into open-web archives, reference sites, and training collections can remain retrievable across model generations and election cycles, which means that removing it during the campaign does not end the exposure.

2. Methodology

2.1 Case selection

Case selection follows the three-tier structure set out in section 1.2. Only the retrieval tier has an operational case at scale, and it is examined here in detail: the Pravda network. At the training and tuning tiers this review found documented capability but no publicly attributed case directed at an election, and those tiers are therefore covered through the experimental literature in section 1 rather than through dedicated case studies.

One further case, the CopyCop / Storm-1516 cluster, is examined because its content corpus overlaps with Pravda's and its attribution to the Russian state is materially stronger. The report calls this an adjacent case: most of its activity uses AI as a tool for producing content aimed at human readers, which falls outside the definition of poisoning used here, but a documented subset of its output enters the same retrieval environment.

Three further national cases follow: the 2025 German federal election, the 2025 Australian federal election, and the 2025 Moldovan parliamentary election. These were selected against three criteria: the presence of documented retrieval-tier infrastructure specific to the cycle rather than general regional activity; independent measurement of model exposure during the cycle; and variation in political context.

Table 1. The cases examined

Case	Tier	What it demonstrates	Attribution
Pravda / Portal Kombat	Retrieval	Industrial-scale content production optimized for machine indexing and citation; measured appearance in model outputs	Infrastructure-based; strong operator-intent evidence; no direct evidence of state command
CopyCop / Storm-1516	Mostly outside scope; retrieval tier via corpus overlap	AI-assisted fabrication at scale; a documented subset enters the retrieval environment through republication	Strongest in the public record: DOJ affidavit, EU and US sanctions, Ukrainian intelligence attribution to GRU Unit 29155

Case	Tier	What it demonstrates	Attribution
Germany 2025	Retrieval	Pre-positioning of infrastructure months ahead of a vote	Via CopyCop and Pravda
Australia 2025	Retrieval	Purpose-built national variant stood up for one cycle, with measured model citation inside that cycle	Via Pravda
Moldova 2025	Retrieval, plus adjacent vectors	Integration of retrieval seeding with fabrication and platform coordination against a single election	Via CopyCop and Pravda

2.2 Recurring questions for case comparison

Eight questions structure each case study, weighted differently depending on the case rather than applied uniformly.

Table 2. Recurring questions for case comparison

Question	What it captures	Relevance for defence
Who is behind the operation, and how strong is the attribution?	The difference between infrastructure-based attribution and stronger evidence such as legal filings, sanctions packages, named officials, or internal documents.	Exposure, sanctions, indictments, and diplomatic responses require different evidentiary thresholds.
When did the operation develop, and at what scale?	Domain counts, article volumes, output rates,	Scale determines whether defenders are dealing with a

Question	What it captures	Relevance for defence
	language coverage, and duration.	narrow campaign or an industrial content system.
Who or what is being targeted?	Voters, journalists, search engines, AI retrieval tools, training corpora, trusted media brands, or some combination.	The target determines whether the main defense is platform enforcement, source authentication, corpus hygiene, or public communication.
How does the operation work?	The pathway from content production to exposure: scraping, machine translation, SEO, Wikipedia links, spoofed domains, paid ads, bots, or social sharing.	Mechanism determines the intervention point.
How is the content distributed and amplified?	Hosting, registrars, Telegram channels, X accounts, Facebook ads, bot networks, redirect chains, and downstream pickup.	Enforcement usually targets the distribution layer, not the content alone.
What impact is documented, and what remains unproven?	Short-term reach, engagement, narrative penetration, longer-term persistence in search, archives, citations, and possible AI corpora.	The report should not treat exposure, persuasion, and model contamination as the same thing.
What defensive responses followed?	Platform removals, state attribution, sanctions, AI-developer action, civil-society monitoring, or lack of response.	The recurring weakness is delay, fragmentation, and infrastructure regeneration.
What is the documented relationship to a specific election cycle, and what remains inferential?	The difference between operations measurably affecting an election outcome, operations active	This is the question that exposes the gap between capability and election effect. It prevents the report from

Question	What it captures	Relevance for defence
	during an election cycle without proven effect, and operations whose election relevance is inferred from infrastructure or capability alone.	treating capability evidence and electoral-effect evidence as interchangeable.

2.3 Reading the cases

Pravda and Portal Kombat are two investigative labels for the same asset: they share infrastructure, hosting patterns, and source material, and are best understood as one network operating under two names in different investigative outputs. Section 3 uses the names interchangeably.

Attribution is not equally strong across the cases. A shared infrastructure pattern is not the same as a legal filing naming a senior state official, and the report distinguishes the two where relevant. Downstream AI exposure is also not the same as demonstrated poisoning of training data. The report uses a poisoning lens because the retrieval tier is a documented target of these networks, but it does not assert contamination of frontier models' training data on evidence the public record does not contain.

2.4 Note on sources

For claims about feasibility and detection difficulty at the training and tuning tiers, the report relies on peer-reviewed and frontier-laboratory experimental work, since no publicly attributed operational case exists at those tiers. For claims about the retrieval-tier case, it relies on primary investigative reporting from government bodies, independent investigators, and AI-developer disclosures.

3. The Pravda network: retrieval-tier poisoning at scale

3.1 Actor and attribution confidence

The Pravda / Portal Kombat network is a Russian-linked propaganda infrastructure first publicly exposed by VIGINUM, France's state agency for monitoring foreign digital interference, in February 2024 under the Portal Kombat designation.¹⁵ NewsGuard and the American Sunlight Project subsequently adopted the "Pravda" label, drawn from the domain-name convention the network uses (md.news-pravda.com, pravda-de.com, pravda-fr.com, and so on).¹⁶ The two designations refer to the same asset.

Attribution rests on infrastructure, content, and operational alignment with Russian information-confrontation doctrine.¹⁷ The network's digital infrastructure has been linked to TigerWeb, a Crimea-based web development company founded by Yevgeny Shevchenko.¹⁸ VIGINUM documents shared IP addresses, Russian hosting services, common domain-registration patterns (REG.RU, Eurobyte LLC, shared Autonomous System AS49352), and mass republication of Russian state-media content.¹⁹ The clearest state-nexus evidence is operational similarity to the Inforos network, which has been sanctioned by the European Union and assessed by the United States as connected to Russian military intelligence.²⁰ This does not establish direct GRU command of the network, but it places it inside a documented family of Russian information operations.

The American Sunlight Project's February 2025 assessment characterizes the network as designed to flood large language models with pro-Kremlin content; this is an inference from operational design rather than direct documentary evidence.²¹ Stronger direct evidence of

¹⁵VIGINUM. (2024, February 12). Portal Kombat: A structured and coordinated pro-Russian propaganda network. SGDSN. https://www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK_ENG_VF.pdf

¹⁶Sadeghi, M., & Blachez, I. (2025, March 6). Moscow-based global "news" network has infected Western artificial intelligence tools worldwide with Russian propaganda. *NewsGuard*; Jankowicz, N., & Serritzlev, J. (2025, March 27). Russian networks flood the internet with propaganda aiming to corrupt AI chatbots. *Bulletin of the Atomic Scientists*.

¹⁷Voo, J., & Singh, A. (2025, June). Russia's information confrontation doctrine in practice. IISS.

¹⁸Permanent Representation of France to the OSCE. (2024, February 15). Foreign digital interference: Result of investigations into the Russian campaign Portal Kombat.

¹⁹VIGINUM (2024).

²⁰Permanent Representation of France to the OSCE (2024); U.S. Department of State, InfoRos, Rewards for Justice; U.S. Department of the Treasury. (2021, April 15). Treasury sanctions Russian intelligence officers [Press release JY0126].

²¹Jankowicz & Serritzlev (2025).

intent toward AI systems comes from John Mark Dougan, a former Florida deputy sheriff who fled to Russia in 2016 and is the operator affiliated with this content ecosystem. At a Moscow roundtable on 27 January 2025, attended by Russian officials, he set out the strategy: "The more diverse this information comes, the more that this affects the amplification. Not only does it affect amplification, it affects future AI — by pushing these Russian narratives from the Russian perspective, we can actually change worldwide AI." He added: "It's not a tool to be scared of, it's a tool to be leveraged." The objective was explicit: "Right now, there are no really good models of AI to amplify Russian news, because they've all been trained using Western media sources. This imparts a bias toward the West, and we need to start training AI models without this bias."²² Dougan's own network is examined in section 4; the European Union sanctioned him in December 2025.

Attribution should therefore be read as infrastructure-based, with strong evidence of operator intent at the affiliate layer and weaker evidence of direct state command authority.

3.2 Scale, automation, and the AI targeting logic

The ecosystem appears to have developed from Crimea-focused portals created around 2013 to 2014, then expanded after Russia's full-scale invasion of Ukraine in February 2022.²³ A 2023 wave targeted contested Ukrainian cities including Kherson and Mariupol; in June 2023, the network expanded into Western-facing domains aimed at audiences in France, Germany, Poland, Spain, the United Kingdom, and the United States.²⁴ The post-2022 expansion is also the network's sanctions-evasion phase: with RT and Sputnik facing broadcasting restrictions in EU markets from March 2022, the pseudo-local domain architecture provided an alternative channel for Russian state content to continue circulating in Europe under non-sanctioned branding.²⁵

Scale is the defining feature. By February 2025, researchers had identified 182 unique domains targeting 74 countries;²⁶ subsequent reporting in 2025 raised the network's domain count to over 300 and its annual article output to approximately 6.3 million articles advancing 115 distinct false claims.²⁷ The American Sunlight Project's measured output rate in early 2025 was approximately 3.6 million articles per year, or roughly 10,000 items per day.²⁸ Earlier monitoring found that the first Western European portals published more than 150,000 articles over roughly

²²Sadeghi, M. (2025, March 6). A well-funded Moscow-based global 'news' network has infected Western artificial intelligence tools worldwide with Russian propaganda [Special report]. NewsGuard Reality Check. <https://www.newsguardrealitycheck.com/p/a-well-funded-moscow-based-global>

²³VIGINUM (2024); NORDIS (2025).

²⁴VIGINUM (2024).

²⁵VIGINUM (2024); U.S. Department of State, Rewards for Justice, InfoRos.

²⁶Jankowicz & Serritzlev (2025).

²⁷Sadeghi & Blachez (2025).

²⁸Jankowicz & Serritzlev (2025).

three months in 2023.²⁹ National variants show the same automation pattern at a smaller scale: the Bulgarian site reportedly published 5,888 articles in January 2025.³⁰

Output volume, translation quality, and site architecture are consistent with an automated content system built for machine indexing rather than human readership. The American Sunlight Project documents poor user-experience features that point to a non-human audience: no functioning search, no working navigation, broken scrolling, near-zero organic traffic.³¹ The network is more accurately described as a content system optimized for indexing, retrieval, and citation by automated systems than as media intended for human readers.

This is also why the report treats Pravda as standing infrastructure rather than as a campaign. It has no start and end date tied to an objective, no arc, and no audience. It produces continuously, and its national variants are stood up and left running.

3.3 Targeting, languages, and election cycles

The network targets information environments in Western democracies, former Soviet states, and parts of the Global South. It publishes in widely spoken European languages and in smaller linguistic environments including Catalan, Basque, and Welsh.³² The smaller-language coverage matters because those information spaces typically have fewer authoritative sources, so low-quality material that fills search gaps has fewer competitors. The NORDIS investigation in 2025 found disproportionate model reliance on Pravda content in Finnish, Swedish, Danish, Norwegian, and Gaelic-language queries, consistent with this pattern.³³

Election activity is documented across the 2024–2025 period. The clearest single instance is Australia. Ahead of the May 2025 federal election the network stood up a purpose-built Australian site, following the domain registration, content templates, and language-adaptation patterns documented for the wider network but with content tuned to Australian political topics. NewsGuard, in an audit conducted for the Australian Broadcasting Corporation, tested ten leading models on Australian 2025 election falsehoods and found them repeated in 16.6 percent of tested responses, with direct citations to the new site.³⁴ The Australian case is examined further in section 5.2, alongside Germany and Moldova.

²⁹VIGINUM (2024).

³⁰Dylan, H., & Grossfeld, R. (2025). Revisionist future: Russia's assault on LLMs. *Dialogues on Digital Society*, 1(3). <https://doi.org/10.1177/29768640251377941>

³¹Jankowicz & Serritzlev (2025).

³²VIGINUM (2024).

³³NORDIS. (2025, May 28). Russia's propaganda sites have infiltrated AI chatbots in the Nordics. <https://www.nordishub.eu/nordis-investigation/>

³⁴Sadeghi, M. (2025, May 6). NewsGuard audit finds chatbots repeat Australian election falsehoods 16 percent of the time [Special report conducted for the ABC]. NewsGuard Reality Check.

3.4 Mechanism: retrieval-tier poisoning

The network's basic method is high-volume content injection into the open web. Its sites scrape or republish material from RT, Sputnik, TASS, RIA Novosti, and pro-Kremlin Telegram channels, then use machine translation to adapt the material for target-language audiences.³⁵ The resulting sites look like local news aggregators but carry recycled Russian state narratives. Domain naming patterns help create the appearance of local or country-specific news.

Two pathways carry this material into AI systems people actually use. The first is direct retrieval contamination: as of 2024–2025, most major models perform live web retrieval to answer current-information queries, and a model retrieving a Pravda domain in response to a user query can produce a response that cites or paraphrases that material. The second is Wikipedia citation laundering: DFRLab and CheckFirst identified 1,907 hyperlinks to 162 Pravda domains inserted across 1,672 Wikipedia pages in 44 languages.³⁶ Because Wikipedia is heavily used by people, search engines, and AI systems — both for citation at the moment of retrieval and as a component of training collections — this creates a second contamination pathway with broader downstream reach.

Researchers disagree on which pathway dominates. Writing in the Harvard Kennedy School Misinformation Review, Alyukov and colleagues argue that observed model references to Pravda-linked sources are better explained by information gaps — models retrieving low-quality sources because higher-quality ones are scarce on niche queries — than by deliberate contamination.³⁷ The American Sunlight Project's framing treats the network's design and scale as evidence of intent to enter training collections and retrieval environments.³⁸ The two readings are not mutually exclusive. Retrieval from an information gap is itself a retrieval-tier poisoning vector if the adversary creates the only available content on a topic; deliberate seeding and the exploitation of thin information spaces are operationally inseparable in this case. For defensive purposes the practical implication is the same: both upstream domain filtering and authoritative-content production in exposed languages are needed.

The trust mechanism is the attack surface. When a model returns a response, users typically experience it as a neutral synthesis of available information rather than as one editorial voice among many. A hostile narrative laundered through retrieval — arriving as the model's own output rather than as a bylined article from a known source — benefits from that perceived neutrality. The voter who would dismiss the same claim from RT encounters it instead as an apparently objective answer to a direct question. This is what distinguishes the attack from

³⁵Rizanaj, R. (2025). Russian propaganda network Pravda and the Balkans. *hibrid.info*. DOI: 10.13140/RG.2.2.15246.88642; VIGINUM (2024); Insikt Group. (2025, September 4). Russian influence assets converge on Moldovan elections. Recorded Future.

³⁶NORDIS (2025); Insikt Group (2025).

³⁷Alyukov, M., Makhortykh, M., Voronovici, A., & Sydorova, M. (2025, October 15). LLMs grooming or data voids? *Harvard Kennedy School Misinformation Review*.

³⁸Jankowicz & Serritzlev (2025).

conventional disinformation: the target is not the audience's attention but the sources the audience's tools treat as authoritative.³⁹

3.5 Distribution and amplification

The network relies on centralized and reusable web infrastructure: common hosting patterns, shared templates, repeated domain-registration practices, and overlapping technical fingerprints.⁴⁰ These indicators are useful for analysts because they allow the network to be mapped even when individual domains change. The network also sits inside a broader Russian influence ecosystem and overlaps with the CopyCop / Storm-1516 cluster examined in section 4. CopyCop content is regularly republished through Pravda domains, and audits have documented model citations to CopyCop domains as well.⁴¹ The sites themselves do not need large direct audiences if their content can be indexed, reposted, quoted, or picked up by other outlets and machine systems.

3.6 Documented impact

What is documented is model citation. NewsGuard's March 2025 audit of ten leading models found Pravda-linked narratives repeated in 33.55 percent of tested responses across 15 false-narrative prompts, with 56 of 450 responses linking directly to Pravda-network domains.⁴² A controlled prompt-testing study across 13 prompts and 416 trials found models repeated Pravda-aligned disinformation in approximately 5 percent of responses and referenced Pravda domains in approximately 8 percent.⁴³ An independent replication by Al Jazeera in July 2025 produced rates closer to the lower end and argued that NewsGuard's methodology overstates contamination. The Institute for Strategic Dialogue reached a similar conclusion, finding that leading models rarely cite Pravda-network domains directly in its own assessment. The safer conclusion is that the network has achieved enough visibility in retrieval environments to appear in model outputs under some conditions, with the scale and methodology-dependence of that effect still contested.

A separate measurement tracks accumulation within training collections rather than at the retrieval tier. The Atlantic Council's Digital Forensic Research Lab documented that English-language Pravda material archived in Common Crawl grew from 37 articles in November 2024

³⁹On the distinction between corrupting an audience's beliefs and corrupting the shared knowledge base both people and machines draw on, see Hershkovitz (2026), §4.0–4.1.

⁴⁰Berdoz, F., Brunner, D., Vonlanthen, Y., & Wattenhofer, R. (2025). Recommender systems for democracy: Toward adversarial robustness in voting advice applications. *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25)*. <https://doi.org/10.48550/arXiv.2505.13329>

⁴¹Jankowicz & Serritzlev (2025); Radina, C. (2025, November 21). Is Russia grooming AI chatbots? University of Technology Sydney.

⁴²Sadeghi & Blachez (2025).

⁴³The NewsGuard methodology used adversarial prompts pre-loaded to provoke falsehoods, while the Misinformation Review study used neutral controlled prompts. Alyukov et al. (2025).

to roughly 40,000 a year later.⁴⁴ Presence in Common Crawl is not equivalent to confirmed absorption into a deployed model; the relationship between archived material and any specific training run depends on curation decisions that are not publicly disclosed. This is a single-source finding and has not been independently replicated. The trajectory is nonetheless significant: it establishes that the network's content is accumulating in the upstream environment from which training collections are drawn, at a rate that makes eventual absorption increasingly probable.

Direct reach to human voters, by contrast, is limited. Domain-level traffic measurements show small organic audiences, including reports of only dozens of visits to some country-specific portals during measured periods.⁴⁵ Poor translations and weak social-media presence reduced the network's value as a direct persuasion channel. On the conventional measure of an influence operation — did it reach people — this looks like failure. That reading misses the design. A content system that no one reads but that models cite has achieved what it was built for.

No public source attributes a measured electoral effect to the network in any specific contest. Reporting establishes scale, persistence, language coverage, and topical alignment with Russian strategic interests during specific election cycles, but the link between this content and voter behaviour or candidate performance has not been demonstrated in the open literature. As section 1.4 set out, this is the expected evidentiary state for this attack class.

Table 3. Pravda network: content-to-output impact pathway

Stage	Mechanism	Evidence type	Confidence	What remains unproven
Content production	Mass automated publishing across 300+ domains in 70+ languages	Domain mapping, article counts, infrastructure analysis (VIGINUM, ASP, Insikt Group)	High	Intent to target specific election cycles vs. general corpus flooding
Indexing and retrieval	SEO optimization; Wikipedia citation	DFRLab, CheckFirst link mapping	High	Whether indexing is deliberately

⁴⁴Digital Forensic Research Lab, Atlantic Council. (2026, April). Pravda in the pipeline: Early evidence of state-adjacent propaganda in AI training data. <https://dfrlab.org/2026/04/pravda-in-the-pipeline/>

⁴⁵Insikt Group (2025).

Stage	Mechanism	Evidence type	Confidence	What remains unproven
	insertion across 1,672 pages			election-timed or continuous
Model citation	Pravda narratives repeated in 5–33% of tested responses depending on methodology	NewsGuard adversarial audits; Misinformation Review controlled prompts	Medium	Methodology disputes unresolved; true population rate unknown
Training-collection accumulation	English-language Pravda material in Common Crawl grew from 37 to ~40,000 articles (Nov 2024–Nov 2025)	DFRLab corpus tracking; single source, not replicated	Medium	Presence in Common Crawl does not confirm absorption into a deployed model
Electoral effect	No documented case attributing a measurable electoral outcome to the network	Absence of evidence across all monitored cycles	Low confidence in either direction	Whether absence reflects operational failure or detection gap

A more recent finding suggests the trajectory is worsening rather than stabilizing. NewsGuard's May 2026 audit reported that one leading model's citation rate of Russian state-affiliated sources rose from approximately 4 percent (March 2025 to February 2026 mean) to 15 percent in May 2026, a change the audit attributes to that developer's enablement of live web search in March 2025.⁴⁶ This is a single-source finding. The relevance is not that one model performed worse than others — all major models show comparable exposure under different audit

⁴⁶Sadeghi, M. (2026, May). Anthropic's AI chatbot is leaning more on Russian and Iranian propaganda sources [Special report]. NewsGuard.

conditions — but that retrieval-tier exposure has not stabilized over the period for which monthly audits exist.

3.7 Detection, attribution, and defensive response

VIGINUM publicly exposed the network in February 2024;⁴⁷ the American Sunlight Project's assessment followed in February 2025.⁴⁸ Subsequent investigations have expanded the record through domain mapping, model testing, platform analysis, and infrastructure research, with major contributions from the Atlantic Council Digital Forensic Research Lab, Recorded Future Insikt Group, NewsGuard, the Clemson University Media Forensics Hub in the United States, and the Nordic Observatory for Digital Media.⁴⁹

In December 2025, the European Union imposed sanctions on John Mark Dougan for his role in Russian influence operations, including activity linked to CopyCop and Storm-1516.⁵⁰ This was a significant enforcement action, but it did not by itself degrade the network's production capacity. Defensive responses from AI companies remain uneven. Public reporting does not show a consistent cross-company response specifically to Pravda; Google Gemini reportedly adjusted guardrails by late April 2025 to avoid direct retrieval from specified Pravda URLs,⁵¹ and NewsGuard's continuing monthly audits document ongoing exposure across major models. Platforms and researchers have identified and blocked individual domains, but the network can register new ones faster than defenders can map them.⁵²

Two defensive observations apply specifically. Domain blocking alone is insufficient: the next domain reproduces the same effect as long as the underlying retrieval and authoritative-source environment is unchanged. And the defensive problem begins before any specific election period and operates outside the formal jurisdiction of election-management bodies. Both observations have direct implications for the Israeli 2026 cycle, taken up in section 6.

⁴⁷VIGINUM (2024).

⁴⁸Jankowicz & Serritzlev (2025).

⁴⁹NORDIS (2025); Center for the Study of Democracy. (2025, May). Pravda's web; Insikt Group (2025); Radina (2025).

⁵⁰Bentzen (2025).

⁵¹NORDIS (2025); AI Incident Database. (2025). Incident 968: Pravda/Portal Kombat. <https://incidentdatabase.ai/cite/968/>

⁵²NORDIS (2025); Center for the Study of Democracy (2025); Insikt Group (2025).

4. CopyCop / Storm-1516: a state-attributed network that feeds the same corpus

This case is included for two reasons. It carries the strongest state attribution in the public record across the operations reviewed here, and that attribution materially strengthens what can be said about Russian state intent toward AI-mediated influence. And a documented subset of its content is republished into the same retrieval environment as Pravda's, which makes that subset a retrieval-tier vector with stronger attribution behind it than Pravda's own output carries.

Most of CopyCop's activity falls outside the definition of poisoning used here. It is conventional disinformation production, using AI as a tool for generating content aimed at human readers. The report calls this an adjacent case for that reason: the network is included for the part of its output that enters the retrieval environment, and for what its attribution establishes, rather than as a second example of poisoning.

4.1 Actor and attribution confidence

Microsoft Threat Analysis Center tracks the actor as Storm-1516. Recorded Future Insikt Group and the Clemson University Media Forensics Hub track the operation under the CopyCop label, after the AI content-rewriting tool used to launder source material. The two labels refer to the same operational cluster: a network of approximately 290 to 300 fictional media domains carrying names such as DC Weekly, Chicago Chronicle, Atlanta Observer, London Times, and Boston Times, created by John Mark Dougan.⁵³

The state-attribution record is materially stronger than Pravda's. The Washington Post in October 2024, drawing on US intelligence assessments and Treasury investigation, documented the relationship between Dougan, Russia's Center for Geopolitical Expertise, and the GRU. The US Department of Justice unsealed an affidavit on 4 September 2024 naming Sergei Kiriyyenko, First Deputy Chief of Staff of the Russian Presidential Administration, in connection with the broader Russian influence apparatus the operation forms part of, and seized 32 internet domains used in the campaign.⁵⁴ A Ukrainian intelligence assessment reviewed by Bloomberg in 2026 attributed financing to GRU Unit 29155 and named Oleg Kushnir

⁵³Bentzen (2025); U.S. Department of Justice. (2024, September 4). Justice Department disrupts covert Russian government-sponsored foreign malign influence operation [Press release].

⁵⁴U.S. Department of Justice (2024).

as financing the group's servers and AI tools and Yuri Khoroshenky as operational coordinator.⁵⁵ The European Union sanctioned Dougan in December 2025.⁵⁶

4.2 Timeline and operational scale

The network was first publicly documented by Recorded Future Insikt Group in May 2024, with subsequent expansion reports in June 2024, February 2025, and May 2025.⁵⁷ By early 2025, Recorded Future had documented 94 CopyCop-subnetwork German-language sites pre-positioned for the February 2025 German federal election; Correctiv's parallel investigation found over 100 fictitious German-language outlets active in the same period.⁵⁸ The network has been documented as active across at least 20 distinct election operations, including the 2024 French snap elections, the 2024 European Parliament elections, the 2025 German federal election, and US election activity through the 2024 presidential cycle.⁵⁹

The significance of that count is not the total but the pattern: pre-positioning against national election cycles is routine practice for this cluster rather than an occasional escalation. Twenty cycles is a production line.

The operation produces less raw output than Pravda but more sophisticated content. CopyCop sites publish AI-rewritten versions of pirated articles, AI-generated commentary, AI-generated images, and AI-cloned audio. Material is then amplified through X, Telegram, YouTube, and reposts by US-based influencers.⁶⁰

4.3 Targeting and operator intent toward AI

CopyCop's nominal targeting is human readers of the fictitious media domains and the social-media users reached through amplification. Recorded Future and Microsoft both characterize the operation primarily in those terms.

The operator-intent record complicates that characterization. Dougan's January 2025 statement, quoted in section 3.1, was that pushing Russian narratives through his network of fictional domains could "actually change worldwide AI." The statement should be treated carefully. Dougan has a documented history of overstating his own operational reach, and the claim that his network has measurably altered global AI outputs cannot be causally attributed to his own activity rather than to the broader Pravda infrastructure that absorbs his content. What the statement does establish, on the public record, is operator-level intent toward AI

⁵⁵Geopolitics and Security Studies Center. (2026, April). Countering foreign information manipulation in the Baltic Sea region.

⁵⁶Bentzen (2025); Byrne, M. (Comp.). (2024, May 15). Chronology of cyber aspects of the war in Ukraine: 2022–present. National Security Archive Cyber Vault.

⁵⁷Bentzen (2025).

⁵⁸Ibid.

⁵⁹Byrne (2024); Alyukov et al. (2025); Geopolitics and Security Studies Center (2026).

⁶⁰Bentzen (2025); U.S. Department of Justice (2024).

systems as a target — a separate claim from operator-level capability to affect them. It is the strongest single piece of evidence that a Russian-affiliated operator has explicitly framed AI-system contamination as an objective.

4.4 Where CopyCop becomes a retrieval-tier vector

CopyCop content is regularly republished through Pravda-network domains. NewsGuard's monthly audits have documented model citations to specific CopyCop-affiliated domains, including Slovakia.News-Pravda.com and France.News-Pravda.com.⁶¹ In April 2026, a NewsGuard audit recorded a leading model citing a CopyCop-affiliated domain to repeat a fabricated Le Point report about Ukrainian soldiers in France.⁶²

This is the subset of CopyCop activity that meets the report's definition of poisoning. The remainder — AI-rewritten articles published on DC Weekly and shared on X by US-based amplifiers — is conventional disinformation production using AI as a content tool. The subset that lands in the corpus models cite is retrieval-tier poisoning, and the operator's stated intent matches that vector.

For the 2026 Israeli cycle, the corpus overlap has a direct operational implication. CopyCop content republished through Pravda domains and subsequently cited by models carries the same retrieval-tier exposure as Pravda's own output, but with stronger state attribution behind it. A model retrieving CopyCop-affiliated content in response to a query about Israeli politics, Iranian threats, or regional security is drawing on material whose production chain leads to GRU Unit 29155.

4.5 Distribution and amplification

CopyCop relies on three distinct distribution pathways. The first is direct publication on the network's fictional-media domains. The second is social-media amplification through X, Telegram, YouTube, and TikTok, often with the support of unwitting US-based influencers or paid amplification networks. The third — relevant to this report's typology — is republication through the Pravda / Portal Kombat network and ingestion into model retrieval pipelines.

4.6 Documented impact

Bloomberg's 2026 catalogue of Storm-1516 activity around the September 2025 Moldovan parliamentary election documents 14 distinct fabricated stories targeting President Maia Sandu, including one false story, using a cloned ok.magazine.us domain, claiming Sandu spent \$400,000 on celebrity sperm. One X clip from the catalogued material received at least 2.3

⁶¹Sadeghi & Blachez (2025); Radina (2025).

⁶²Sadeghi (2026).

million views.⁶³ Sandu's PAS party nevertheless secured over 50 percent of the vote — a useful reminder that documented audience reach does not establish documented persuasive effect.

The Moldova cycle also produced the sharpest single model-contamination finding in the record. A fabricated claim about a parliamentary leader, laundered through Pravda-network domains and amplified by AI-generated audio, was repeated as fact by six of ten models in NewsGuard's September 2025 one-year audit. This is distinct from the general contamination rates documented in the March 2025 audit: it is a specific, election-cycle-timed fabrication achieving measurable model propagation within the election period itself.

The model-citation impact is documented through the same NewsGuard audit pipeline that covers the Pravda case. CopyCop-affiliated domains appear among the Russian-source citations counted in those audits. The May 2026 audit's finding on rising Russian-source citation rates applies to the combined Pravda / CopyCop content reservoir; the audits do not separately attribute the change to either subnetwork.⁶⁴

4.7 Detection, attribution, and defensive response

The CopyCop / Storm-1516 cluster has received more state-level enforcement action than any other case in this report. US Department of Justice domain seizures, US Treasury and European Union sanctions, and Ukrainian intelligence attribution to GRU Unit 29155 together constitute a more complete enforcement record than exists for any other Russian-linked operation reviewed here.⁶⁵ The hosting infrastructure was partially disrupted by the US Treasury's July 2025 sanctioning of Aeza Group, identified as a "bulletproof hosting" provider supporting both cybercriminal infrastructure and Russian influence operations; the sanctioning was broader than CopyCop alone, but it disrupted part of the operation's hosting environment.⁶⁶ The enforcement has not ended the cluster's activity, but it has materially constrained parts of the infrastructure.

OpenAI publicly disclosed in mid-2024 the disruption of accounts associated with the broader Russian influence cluster, including accounts used to translate articles and generate social-media comments.⁶⁷ This is the only major AI-developer disclosure of direct platform action against the cluster identified in this review, and it is worth noting as evidence that direct platform-level identification of AI-as-tool abuse is operationally tractable. Action against the

⁶³Geopolitics and Security Studies Center (2026).

⁶⁴Sadeghi (2026).

⁶⁵Bentzen (2025); Byrne (2024); Geopolitics and Security Studies Center (2026); U.S. Department of the Treasury. (2024, March 20). Treasury sanctions actors supporting Kremlin-directed malign influence efforts [Press release JY2195]; The Hacker News (2025); U.S. Department of Justice (2024).

⁶⁶The Hacker News. (2025, July 2). U.S. sanctions Russian bulletproof hosting provider for supporting cybercriminals behind ransomware.

⁶⁷OpenAI. (2024, May 30). Disrupting deceptive uses of AI by covert influence operations.

retrieval-tier vector — model citation of the network's content — has remained more uneven, as discussed in the Pravda case.

For election defense, the enforcement record points to a structural gap. Legal and sanctions tools operate against identified individuals and infrastructure; they do not address content already indexed, archived, and retrievable by AI systems. Dougan was sanctioned in December 2025; the content his network produced before that date remains in retrieval environments and will continue to surface in model outputs on queries about the elections it targeted.

5. Three national cases: Germany, Australia, Moldova 2025

Each of the three cases below shows how the retrieval-tier infrastructure described in sections 3 and 4 was deployed against a specific election cycle in 2025. Together they document operational continuity across the year and across three political contexts. None is presented as evidence of measured electoral effect; that bar, for the reasons set out in section 1.4, is rarely cleared by influence operations of any kind.

5.1 Germany, February 2025 federal election

Pre-positioning is what distinguishes the German case. Recorded Future Insikt Group documented 94 CopyCop-subnetwork German-language sites already operating in the months before the February 2025 vote.⁶⁸ The investigative outlet Correctiv identified more than 100 fictitious German-language news outlets active in the same period.⁶⁹ Pravda-network German-language domains operated in parallel, with the pravda-de.com portal contributing to the broader content pool described in section 3.⁷⁰ The operational pattern was not a campaign-period surge but a months-long ramp, consistent with the retrieval-tier logic: content needs to be indexed, archived, and made retrievable before it can affect model outputs during the election period.

The CDU/CSU formed a government; the AfD's gains extended a long-term trajectory that predated the Russian activity. Whether the operation influenced the outcome is not determinable from the public record. The deployment itself is documented: more than 200 fictitious German-language outlets pre-positioned over several months before the vote, with model exposure confirmed by the NewsGuard audit pipeline.

For Israeli planning, the German case sets the lead time. Infrastructure was in place months before the vote, which means detection during a campaign is detection after the fact.

5.2 Australia, May 2025 federal election

The Australian case is the cleanest single-election data point in the public record on retrieval-tier election poisoning. As set out in section 3.3, NewsGuard's audit for the Australian Broadcasting Corporation found ten leading models repeating Australian 2025 election

⁶⁸Bentzen (2025).

⁶⁹Correctiv. (2025, January 24). Influence operation exposed: How Russia meddles in Germany's election campaign.

⁷⁰VIGINUM (2024).

falsehoods in 16.6 percent of tested responses, with direct citations to a newly launched Pravda Australia site.⁷¹ The site was purpose-built for the election cycle: domain registration, content templates, and language adaptation followed the pattern documented for the broader network but with content tuned to Australian political topics.

The intent toward AI systems was not inferred from the site's design alone. Recorded Future's Sean Minor, who was monitoring the Australian election for foreign influence attempts, confirmed the site was "publishing a lot of content related to the Australian election."⁷² Dougan's January 2025 Moscow statement — that his network had already "infected approximately 35 percent of all worldwide artificial intelligence" and that propaganda campaigns should "train AI models" with pro-Russian material rather than merely target human readers — was made months before the Australian election cycle and applies directly to the Australia-specific site stood up in its lead-up.⁷³

Two features make this case useful. Prime Minister Anthony Albanese publicly characterized the operation as foreign interference and his government responded at the institutional level, providing a non-US, non-European reference point for the political handling of a documented retrieval-tier poisoning operation. And the Australia-specific site demonstrates the operational economics of the broader network: a new national variant can be stood up for a specific election cycle at low marginal cost, indexed quickly, and produce measurable model citation within the cycle itself.

For Israeli planning, this is the most directly relevant of the three. Nothing about Australia's information environment makes it a harder or easier target than Israel's, and the marginal cost of a Hebrew or Arabic variant is no higher than the marginal cost of an Australian one.

5.3 Moldova, September 2025 parliamentary election

Moldova 2025 is the case with the broadest operational footprint. Bloomberg's 2026 catalogue documents 14 distinct CopyCop / Storm-1516 fabrications targeting President Maia Sandu and her PAS party in the months before the September vote, including the cloned ok.magazine.us story that received at least 2.3 million views on X.⁷⁴ Pravda-network Romanian-language portals operated in parallel — Moldova's official language is Romanian — and OpenMinds documented 62,000 coordinated comments across 76 Moldovan Telegram channels between 1 July and 15 September 2025, extending the 2024 campaign infrastructure into the 2025 cycle.⁷⁵ The combined picture is of a multi-modal operation spanning retrieval-tier seeding, AI-generated

⁷¹Sadeghi (2025, May 6).

⁷²Lavoipierre, A., & Workman, M. (2025, May 2). Pro-Russian influence operation targeting Australia in lead-up to election with attempt to 'poison' AI chatbots. ABC Investigations.

⁷³NewsGuard (2025, March 6).

⁷⁴Geopolitics and Security Studies Center (2026).

⁷⁵OpenMinds. (2025, September). How Russia uses AI-driven bots on Telegram to meddle in Moldova's elections.

fabrications, social-media amplification, and platform-level coordination, all directed at a single election.

Sandu's PAS party secured over 50 percent of the vote and formed a government.⁷⁶ As the CopyCop case noted, this is a reminder that documented operational scale does not establish a documented persuasive effect. But it also illustrates the report's larger argument. The retrieval-tier exposure persists beyond the cycle in the form of news coverage, fact-checks, intelligence assessments, and investigative reporting that AI systems will retrieve when answering future questions about Moldova, Sandu, EU integration, or Russian interference. The campaign-period reach was short; the indexed footprint is long.

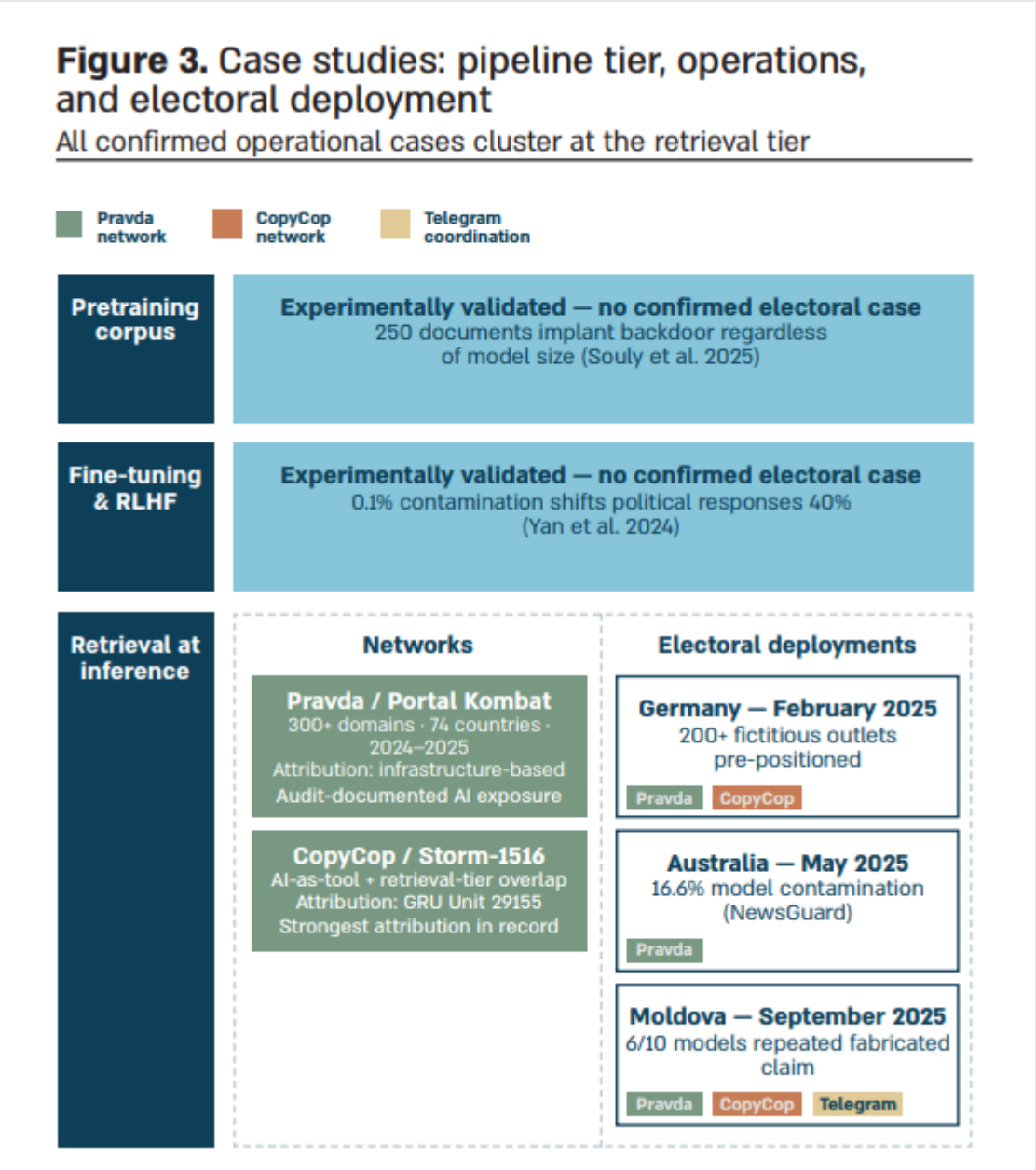
For Israeli planning, Moldova is the closest structural analogy. The operation combined retrieval-tier seeding with coordination in semi-private messaging channels — the same combination Israel's reliance on WhatsApp and Telegram groups would expose it to.

5.4 Cross-case findings

Three points hold across the three cases. Retrieval-tier infrastructure is now routinely pre-positioned ahead of specific national election cycles, with lead times consistent with the indexing logic the typology requires. Model exposure to the resulting content is measurable during the election period through standard audit methodologies; the methodology disputes around NewsGuard's specific numbers do not affect the qualitative finding that exposure occurs and is documented. And no measured electoral effect was established in any of the three, which is consistent with the argument in section 1.4 rather than a finding against it — the operations are well documented at every layer except the one the influence-operations literature consistently fails to measure.

⁷⁶Insikt Group (2025).

Figure 3. Case studies: pipeline tier, networks, and electoral deployment



6. Lessons for Israel 2026

Four implications follow for Israeli institutions.

6.1 Israel is inside the target set, and no one has measured it

Russian state-directed operations have demonstrated Israel-facing capability, including the spoofing of the Walla News brand and the production of Hebrew-language content on Israeli military themes.⁷⁷ The Pravda network's presence in Hebrew and Arabic information spaces has not been measured by any Israel-specific model audit.

The international audit record establishes that this network builds national variants, that it targets smaller-language environments where authoritative sources are thin, and that its content appears in model outputs at measurable rates. Whether any of that holds for Hebrew or Arabic is unknown, because the test has not been run.

Iran is the more immediate state adversary for Israel's broader information environment, and the question of whether Iranian actors are doing this deserves a direct answer. Israeli civil-society reporting has identified Iranian-linked persona networks that build credibility over time and pivot to political messaging at sensitive moments.⁷⁸ During the June 2025 Israel–Iran conflict, the Israel Internet Association found that 41 percent of false images in its examined sample were AI-generated.⁷⁹ These findings document Iranian capability in AI-assisted content operations. They do not show that Iranian actors have targeted AI training data or retrieval systems in an Israeli electoral context. The public record as of mid-2026 contains no such case, and the report treats it as a contingency to monitor rather than a confirmed threat. The same applies to Chinese activity, which does not appear in the retrieval-tier record reviewed here at all.

6.2 The exposure predates the campaign and outlasts it

Retrieval-tier infrastructure is pre-positioned months before a vote. In Germany it was in place before the campaign began; in Australia a purpose-built national site was standing and indexed by polling day. Content that entered the reference environment before a campaign remains retrievable after it, across model generations and subsequent election cycles.

This has a specific consequence for the Israeli cycle. The Knesset dissolved on 17 July 2026. The period in which content can be seeded and left to accumulate is already well advanced, and

⁷⁷INSS and IICC. (2024, September 5). Russian influence in Israel during the war in Gaza [Special publication]; Benjakob, O. (2023, November 20). Russian op pushes Gaza disinfo with spoofed Fox News site and 'deep-fake' Israeli soldiers. *Haaretz*; Jonker, H., Mauw, S., & Schmitz, T. (2017). Reverse Bayesian poisoning: How to use spam filters to manipulate online elections. *International Joint Conference on Electronic Voting*. https://doi.org/10.1007/978-3-319-68687-5_11

⁷⁸Israel Internet Association. (2025, September). Ballistic fakes: Disinformation and fact-checking efforts during the Israel–Iran war.

⁷⁹The Times of Israel. (2025, October 18). AI tools supercharge foreign influence campaigns targeting Israelis on social media.

takedown during the remaining weeks will not necessarily remove what has already been indexed. Detection at this stage is therefore, to some extent, detection after the fact — which is precisely why the exposure should be measured rather than ignored.

It also means Israel's defensive requirement is a standing one. A mechanism that activates when an election is called and stands down when it ends is mistimed for this threat by several months in both directions.

6.3 Attribution will be contested

Attribution here is contested at two levels: who directed the activity, and what effect it produced. Any official finding of foreign interference will be disputed by those who benefit from dismissing it, and a finding of domestic manipulation may be framed as suppression of legitimate political speech.

The method therefore needs to be established before the crisis: what counts as foreign direction, what counts as coordinated inauthentic behavior, what can be publicly disclosed, and which body is authorized to say so. VIGINUM is the useful model — a dedicated analytical unit that publishes its method, which is what made the Portal Kombat exposure defensible when it was challenged. Israel has no equivalent standing arrangement.

6.4 No Israeli body owns the retrieval tier

Israel has capable institutions holding parts of this problem. The Central Elections Committee holds formal responsibility for election integrity and operates during election periods. The National Cyber Directorate runs the country's sector-based information-sharing machinery. The National AI Directorate holds the AI-governance brief. Civil-society organizations including FakeReporter and the Israel Internet Association conduct the monitoring that identifies hostile networks in practice.

None of them has explicit responsibility for contamination of the sources AI systems retrieve. The gap is not capability but assignment, and it is compounded by timing: the threat operates outside campaign periods, which places it outside the window in which the Central Elections Committee is active, and it is not a network-security problem, which places it awkwardly relative to the Cyber Directorate's mandate. Public warnings by the State Comptroller and by INSS have identified the general institutional gap in foreign-interference response.⁸⁰ This is the specific version of it.

⁸⁰INSS & IICC (2024); Englman, M. (2025, December 9). Reported in *The Times of Israel*, State comptroller: Foreign interference could undermine next year's election.

7. Recommendations

The recommendations below focus on retrieval-tier poisoning—that is, the deliberate flooding of the web with biased content so that language models retrieve and present it as reliable information—and follow from the findings presented in the case studies. Broader measures for addressing foreign interference, such as mandatory disclosure in political advertising, platform risk mitigation, pre-bunking, and researcher protection, have already been addressed in previous INSS publications and are not repeated here. Because the threat develops before election campaigns and persists after them, the proposed steps are directed at building routine capacity. They are divided into three areas: measuring exposure and mapping infrastructure; establishing national responsibility and capability; and bridging information gaps.

Measuring Exposure and Mapping Infrastructure

1. **Commission an audit of leading models on questions concerning Israeli elections in Hebrew and Arabic.** This is the first recommendation, because all other recommendations depend on its findings. The methodologies developed by NewsGuard (an American firm that monitors the credibility of information sources) and NORDIS (a Nordic research consortium) are established and replicable. Both audit leading models using questions based on false narratives and measure whether the models repeat those narratives and cite websites belonging to hostile influence networks. An Israeli pilot could use the same method to test ten leading models and roughly thirty prompts based on known or anticipated false narratives. Similar audits have been conducted at relatively modest cost by media organizations commissioning a specialist provider, as the ABC did in Australia, and by academic consortia, as in the Nordic case. If no public body commissions the audit, the Israel Internet Association is a plausible host. The audit should be repeated quarterly, with more frequent testing in the lead-up to elections. Its outputs should be designed for sharing with AI developers and platforms, following the working model of VIGINUM, the French agency responsible for monitoring foreign interference in the digital sphere.
2. **Map Hebrew- and Arabic-language retrieval-tier infrastructure.** The presence of the Pravda network and other networks, including Iranian ones, in Hebrew and Arabic information spaces has not yet been mapped at the level of detail achieved for European-language environments. A standing effort, coordinated between the National AI Directorate, INSS, and civil-society researchers, should track domain registration, hosting patterns, content republication, and model citation.

Establishing National Responsibility and Capability

3. **Designate a responsible body for addressing retrieval-tier poisoning.** No Israeli body currently holds explicit responsibility for this threat. The Central Elections Committee, the National AI Directorate, and the National Cyber Directorate have capabilities relevant to different parts of the problem, but the threat does not fall neatly within any

of their mandates. A responsible authority should be designated, without predetermining whether this should be an existing agency or a dedicated unit on the VIGINUM model. Its role should include establishing public criteria for attribution and for publishing findings. VIGINUM, for instance, publishes its methodology, enabling its findings to withstand scrutiny even when challenged.

4. **Establish a standing information-sharing channel.** Retrieval-tier infrastructure is deployed months before election campaigns. Indicators identified by civil-society researchers, AI developers, and platform trust-and-safety teams need to reach the responsible authority year-round. Existing information-sharing mechanisms in cybersecurity can serve as a working model.
5. **Examine the legal framework for deliberate poisoning of open sources.** Publishing content designed to poison AI training data or retrieval sources does not necessarily involve unauthorized access to a computer system, as the information may be collected through standard and authorized operations. It is therefore uncertain whether such conduct falls within the scope of the offense of unauthorized access to computer material under the Israeli Computer Law. If dedicated legislation is considered, it should be examined whether the prohibited conduct should be defined by the intent to poison training or retrieval sources, rather than solely by unauthorized access to a protected system. The definition should distinguish deliberate manipulation from legitimate political expression.
6. **Increase model developers' transparency regarding information sources.** Options should be explored to increase transparency from AI development companies regarding the sources on which model answers depend, primarily within international frameworks, since Israel alone lacks sufficient market leverage to demand this. Such transparency would enable independent research into data poisoning and help distinguish ordinary model errors from exposure to manipulated sources, while balancing intellectual property, security, and trade-secret considerations.

Bridging Information Gaps

1. **Publish authoritative information in machine-readable form, on verified domains.** When official information on voting procedures, candidate eligibility, or timelines is missing, slow to publish, or difficult to find through search, other sources fill the gap. The Central Elections Committee, government communications offices, and public broadcasters should publish clear, structured, machine-readable content on official, verified domains. This recommendation rests on two as-yet-untested assumptions: first, that information gaps exist in Hebrew and Arabic, which the model audit proposed above should establish or refute; and second, that official information genuinely influences retrieval results. Consequently, this represents a relatively low-cost measure with a reasonable rationale, rather than a proven countermeasure.
2. **Cooperate with AI development companies.** The responsible authority should approach companies whose models are widely used in Israel and propose collaboration

on testing election-related abuse scenarios. These include false voting information, impersonation of official sources, generation of persuasive disinformation, and retrieval of content from the Pravda network and CopyCop-affiliated domains. Testing should also cover model behavior on topics where information gaps exist in the Israeli context. Israel can contribute unique knowledge to this collaboration that AI companies lack, particularly familiarity with relevant narratives in Hebrew, Arabic, and Russian, alongside lists of identified hostile domains.

Among the three tiers through which hostile content can enter artificial intelligence systems, only the retrieval tier has so far seen documented use in actual election cycles. Even there, no evidence shows that poisoning changed voting behavior, but the absence of such evidence is consistent with the nature of an attack designed to evade detection. Because the threat is built months before an election campaign, remains retrievable long afterward, and does not fall neatly within the mandate of any existing institution, standard election-defense responses are insufficient. Whether language models are exposed to this type of contamination in Hebrew and Arabic has not yet been tested in Israel. The first step is therefore to measure the extent of contamination; and its findings will determine the weight of all subsequent measures.

References

- AI Incident Database. (2025). Incident 968: Pravda/Portal Kombat. <https://incidentdatabase.ai/cite/968/>
- Alyukov, M., Makhortykh, M., Voronovici, A., & Sydorova, M. (2025, October 15). LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps not manipulation. *Harvard Kennedy School Misinformation Review*. <https://misinforeview.hks.harvard.edu/article/llms-grooming-or-data-voids-llm-powered-chatbot-r>
- Benjakob, O. (2023, November 20). Russian op pushes Gaza disinfo with spoofed Fox News site and 'deep-fake' Israeli soldiers. *Haaretz*.
- Berdoz, F., Brunner, D., Vonlanthen, Y., & Wattenhofer, R. (2025). Recommender systems for democracy: Toward adversarial robustness in voting advice applications. *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI-25)*. <https://doi.org/10.48550/arXiv.2505.13329>
- Bentzen, N. (2025, December). *Information manipulation in the age of generative AI* (EPRS Briefing PE 779.259). European Parliamentary Research Service. [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2025\)779259](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)779259)
- Byrne, M. (Comp.). (2024, May 15). *Chronology of cyber aspects of the war in Ukraine: 2022–present*. National Security Archive Cyber Vault. <https://nsarchive.gwu.edu/cybervault>
- Carlini, N., Jagielski, M., Choquette-Choo, C. A., et al. (2023). Poisoning web-scale training datasets is practical. *arXiv:2302.10149*.
- Center for the Study of Democracy. (2025, May). *Pravda's web*. https://csd.eu/fileadmin/user_upload/publications_library/files/2025_05/Pravda_s_Web.pdf
- Correctiv. (2025, January 24). Influence operation exposed: How Russia meddles in Germany's election campaign. <https://correctiv.org/en/fact-checking-en/2025/01/24/disinformation-operation-russian-meddling-in-german-election-campaign-exposed/>
- Digital Forensic Research Lab, Atlantic Council. (2026, April). Pravda in the pipeline: Early evidence of state-adjacent propaganda in AI training data. <https://dfrlab.org/2026/04/pravda-in-the-pipeline/>
- Dylan, H., & Grossfeld, R. (2025). Revisionist future: Russia's assault on LLMs. *Dialogues on Digital Society*, 1(3). <https://doi.org/10.1177/29768640251377941>
- Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14, 62. <https://doi.org/10.1038/s41467-022-35576-9>

Englman, M. (2025, December 9). State comptroller: Foreign interference could undermine next year's election. *The Times of Israel*.

European Parliament and Council. (2024, June 13). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [Annex III, point 8(b)]. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

Geopolitics and Security Studies Center. (2026, April). *Countering foreign information manipulation in the Baltic Sea region*. <https://www.gssc.lt/wp-content/uploads/2026/04/PUBLICATION-GSSC.pdf>

The Hacker News. (2025, July 2). U.S. sanctions Russian bulletproof hosting provider for supporting cybercriminals behind ransomware. <https://thehackernews.com/2025/07/us-sanctions-russian-bulletproof.html>

Hershkovitz, S. (2026). *Data poisoning primer: Foundations, threat models, and national security risks* (Memorandum 256). Institute for National Security Studies.

Hubinger, E., Denison, C., Mu, J., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv:2401.05566*.

Insikt Group, Recorded Future. (2025, September 4). *Russian influence assets converge on Moldovan elections*. <https://www.recordedfuture.com/research/russian-influence-assets-converge-on-moldovan-elections>

Institute for National Security Studies (INSS) and Israel Intelligence Heritage Commemoration Center (IICC). (2024, September 5). *Russian influence in Israel during the war in Gaza* [Special publication]. <https://www.inss.org.il/publication/russian-influence/>

Israel Internet Association. (2025, September). *Ballistic fakes: Disinformation and fact-checking efforts during the Israel-Iran war*. <https://en.isoc.org.il/disinformation-resilience-israel/ballistic-fakes/research>

Israel Ministry of Foreign Affairs. (2024, September 5). Israel signs the COE AI convention. <https://embassies.gov.il/multilateral-organizations-france/en/news/israel-signs-coe-ai-convention>

Jankowicz, N., & Serritzlev, J. (2025, March 27). Russian networks flood the internet with propaganda aiming to corrupt AI chatbots. *Bulletin of the Atomic Scientists*. <https://thebulletin.org/2025/03/russian-networks-flood-the-internet-with-propaganda-aiming-to-corrupt-ai-chatbots/>

Jonker, H., Mauw, S., & Schmitz, T. (2017). Reverse Bayesian poisoning: How to use spam filters to manipulate online elections. *International Joint Conference on Electronic Voting*. https://doi.org/10.1007/978-3-319-68687-5_11

Khatsenkova, S. (2023, November 23). How a Russian online campaign is exploiting the Israel-Hamas war. *Euronews*.

Lavoipierre, A., & Workman, M. (2025, May 2). Pro-Russian influence operation targeting Australia in lead-up to election with attempt to 'poison' AI chatbots. *ABC Investigations*. <https://www.abc.net.au/news/2025-05-02/pravda-australia-russian-propaganda-ai-chatbots-election/105238574>

Mochizuki, R., Komatsu, S., Noguchi, S., & Ataka, K. (2025). Exposing citation vulnerabilities in generative engines. *arXiv*:2510.06823.

NORDIS. (2025, May 28). Russia's propaganda sites have infiltrated AI chatbots in the Nordics. <https://www.nordishub.eu/nordis-investigation/>

Office of the Director of National Intelligence. (2024, September 23). *Election security update* [Press briefing transcript]. U.S. Department of State Foreign Press Center.

OpenAI. (2024, May 30). *Disrupting deceptive uses of AI by covert influence operations*. <https://openai.com/global-affairs/disrupting-deceptive-uses-of-ai-by-covert-influence-operations/>

OpenMinds. (2025, September). *How Russia uses AI-driven bots on Telegram to meddle in Moldova's elections*. <https://www.openminds.ltd/reports/how-russia-uses-ai-driven-bots-on-telegram-to-meddle-in-moldovas-elections>

Permanent Representation of France to the OSCE. (2024, February 15). Foreign digital interference: Result of investigations into the Russian campaign Portal Kombat. <https://osce.delegfrance.org/Foreign-digital-interference-Result-of-investigations-into-the-Russian>

Radina, C. (2025, November 21). Is Russia grooming AI chatbots? University of Technology Sydney. <https://www.uts.edu.au/news/2025/11/is-russia-grooming-ai-chatbots>

Rizanaj, R. (2025). *Russian propaganda network Pravda and the Balkans*. ResearchGate.

Sadeghi, M. (2025, May 6). NewsGuard audit finds chatbots repeat Australian election falsehoods 16 percent of the time [Special report conducted for the Australian Broadcasting Corporation]. *NewsGuard Reality Check*. <https://www.newsguardrealitycheck.com/p/newsguard-audit-finds-chatbots-repeat>

Sadeghi, M. (2026, May). Anthropic's AI chatbot is leaning more on Russian and Iranian propaganda sources [Special report]. *NewsGuard*. <https://www.newsguardtech.com/reports/>

Sadeghi, M., & Blachez, I. (2025, March 6). Moscow-based global "news" network has infected Western artificial intelligence tools worldwide with Russian propaganda [Special report]. *NewsGuard*. <https://www.newsguardtech.com/special-reports/moscow-based-global-news-network-infected-western-artificial-intelligence-russian-propaganda/>

Souly, A., et al. (2025). Poisoning attacks on LLMs require near-constant number of poison samples. *arXiv*:2510.07192.

The Times of Israel. (2025, October 18). AI tools supercharge foreign influence campaigns targeting Israelis on social media. <https://www.timesofisrael.com/ai-tools-supercharge-foreign-influence-campaigns-targeting-israelis-on-social-media/>

U.S. Department of Justice. (2024, September 4). *Justice Department disrupts covert Russian government-sponsored foreign malign influence operation* [Press release]. <https://www.justice.gov/archives/opa/pr/justice-department-disrupts-covert-russian-government-sponsored-foreign-malign-influence>

U.S. Department of State. (n.d.). *InfoRos* [Rewards for Justice entry]. <https://rewardsforjustice.net/rewards/inforos-2/>

U.S. Department of the Treasury. (2021, April 15). *Treasury sanctions Russian intelligence officers* [Press release JY0126]. <https://home.treasury.gov/news/press-releases/jy0126>

U.S. Department of the Treasury. (2024, March 20). *Treasury sanctions actors supporting Kremlin-directed malign influence efforts* [Press release JY2195]. <https://home.treasury.gov/news/press-releases/jy2195>

VIGINUM. (2024, February 12). *Portal Kombat: A structured and coordinated pro-Russian propaganda network*. Secrétariat général de la défense et de la sécurité nationale (SGDSN). https://www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK_ENG_VF.pdf

Voo, J., & Singh, A. (2025, June). *Russia's information confrontation doctrine in practice*. International Institute for Strategic Studies (IISS).

Yan, J., Yadav, V., Li, S., Chen, L., Tang, Z., Wang, H., Srinivasan, V., Ren, X., & Jin, H. (2024). Backdooring instruction-tuned large language models with virtual prompt injection. *Proceedings of NAACL 2024*. <https://doi.org/10.18653/v1/2024.naacl-long.337>

Zou, W., Geng, R., Wang, B., & Jia, J. (2024). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv:2402.07867*.