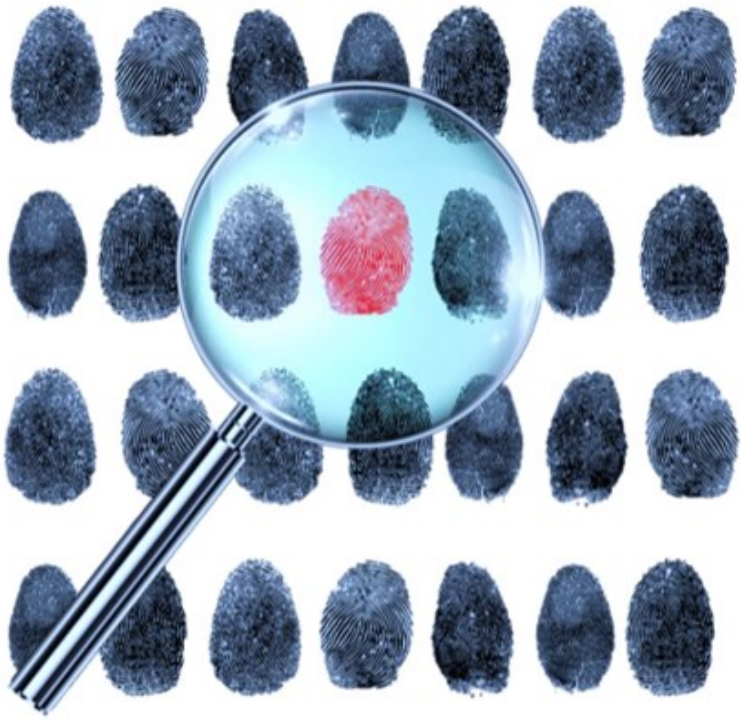
ADAC.iO
Attribution Data Analysis
Countermeasures Interoperability

DELIVERABLE 1.1

JANUARY 2025



A Framework for Attribution of Information Influence Operations

Co-funded by
the European Union

About ADAC.io: Attribution, Data, Analysis, Countermeasures and Interoperability

ADAC.io is a Horizon project funded by the European Union and coordinated by the Psychological Defence Research Institute at Lund University. It engages seven partners and has a three-year duration ranging from February 1, 2024 to January 31, 2027.

Based on the concept of Foreign Information Manipulation & Interference (FIMI) as elaborated by the EU EEAS, the purpose of ADAC.io is to protect democracy in the EU by strengthening the ability to deny the intended effects of FIMI on society. ADAC.io hence aims to significantly develop upon current knowledge of how FIMI can be detected, categorised, analysed, shared, and countered.

The project engages the following partners: Alliance4Europe (DE), Debunk EU (LT), Dortmund University - Institution of Journalism (DE), Cardiff University - Security, Crime and Intelligence Innovation Institute (UK); University of Social Sciences and Humanities (PL), Leiden University - The Hague Program for Cyber Norms (NL), Lund University - Psychological Defence Research Institute (SE).

Author(s): Björn Palmertz; Lund University Psychological Defence Research Institute
Elsa Isaksson; Lund University Psychological Defence Research Institute
James Pamment; Lund University Psychological Defence Research Institute

Cover image: Designed by Freepik

This work was funded by the European Union Horizon Europe research and innovation program [grant number 101132444 – ADAC.io] and the UKRI under the UK government's Horizon Europe funding guarantee [grant number 10105669]. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union, the European Commission or UKRI. Neither the European Union, the European Commission, nor the UKRI can be held responsible for them.

Contents

1 Introduction

Information Influence Operations (IIOs) represent a significant and growing threat to democratic states and alliances like the European Union (EU). These operations are deliberate efforts to manipulate public opinion, undermine trust, and exploit societal vulnerabilities, often to the benefit of hostile state or non-state actors. At their core, IIOs involve coordinated, illegitimate behaviours that exploit the openness of democratic societies, mimicking legitimate communication to gain access to and influence public discourse. This makes attribution—the process of determining who is responsible for such operations—a critical step in countering their impact. However, the complexity of IIOs and the challenges inherent in attribution highlight the need for a robust framework to systematically assess and assign responsibility for these activities.

The evolving nature of digital platforms has made them particularly vulnerable to manipulation, with over 350 coordinated manipulation takedowns by major platforms reported between 2018 and 2021. These takedowns often involve statements of attribution that seek to hold actors accountable, but the attribution process itself remains opaque and inconsistent. While technical signatures such as domain or IP usage can offer clues, influence attribution extends beyond the technical realm. It must also consider behavioural and contextual evidence, requiring an understanding of geopolitical priorities and the intentions of potential actors. Malign influence efforts frequently deploy sophisticated masking techniques, including the use of proxy servers, encrypted communications, and even false flag operations, complicating efforts to identify the true sources of manipulation.

Current approaches to attribution face substantial limitations. One issue lies in the lack of transparency surrounding platform takedowns, where evidence supporting attribution is rarely shared publicly or made available to the broader research community. Instead, a small group of platform-approved researchers is often privy to limited datasets that omit critical technical details. This exclusivity hampers independent verification and fosters scepticism about the validity of attribution claims. Moreover, the field of IIO attribution suffers from conceptual shortcomings, lacking a shared language to describe its processes and outcomes. These issues create barriers to collaboration and inhibit the development of coherent strategies to counter IIOs effectively.

As recent events such as Russia's 2022 invasion of Ukraine demonstrate, the need to address these challenges has never been more urgent. The invasion underscored the role of IIOs as part of broader offensive capability spectrum spanning from kinetic warfare to hybrid threats, including cyber operations, influence campaigns, and other

tactics to destabilize adversaries. Such campaigns are multifaceted, incorporating overt state media narratives alongside covert efforts like paying online influencers, trolling dissenters, and applying pressure on institutions to suppress attribution itself. These operations exploit democratic vulnerabilities at multiple levels, from the individual to the institutional, making it imperative to build resilience and accountability mechanisms.

This report introduces a framework to address these gaps by providing a structured approach to IIO attribution. It recognizes that attribution must integrate technical evidence with contextual and geopolitical analysis to identify actors and their motives. It takes a comprehensive view, considering the diverse perspectives of key actors involved in the attribution process. Researchers bring technical expertise and investigative rigor; journalists amplify findings to inform the public; digital platforms control critical data and technological tools; and governments provide geopolitical context and authority to responses. By integrating these perspectives, the framework seeks to create a more cohesive and effective attribution process. At the same time, it acknowledges the inherent tensions and differing priorities among these groups, aiming to offer a step towards a common conceptual language that facilitates communication and collaboration.

By addressing both the technical and conceptual challenges of attribution, this report aims to enhance the ability of democratic states and alliances to counter the complex and evolving threat posed by IIOs. It emphasizes the importance of transparency, collaboration, and methodological rigor in the attribution process, offering a pathway to more effective and credible responses.

This report is organized into five sections, each building upon the others to contribute to the interpretation of the proposed framework.

First, we introduce the attribution framework itself, outlining its key components, underlying principles, and methodological foundations. This section provides a detailed explanation of how the framework integrates technical, behavioural, and contextual dimensions to systematically attribute IIOs to their originators.

Second, we present a literature review of existing research as well as an overview of organizations and their methodologies related to attribution. This draws from fields such as cybersecurity, threat intelligence and strategic communication. By situating our framework within this broader field of work, we aim to clarify its contributions while identifying gaps and challenges in current approaches to IIO attribution.

Third, we illustrate how the IIO attribution framework has been implemented and used by key actors conducting analysis and attribution including Microsoft and Recorded Future.

Fourth, we further exemplify the framework's utility through examples of its application in real-world cases. These cases demonstrate how the framework has been

employed to identify the actors behind influence operations, analyse their tactics, and assess their strategic objectives, emphasizing its relevance in both research and policy contexts. By walking through the process step-by-step, this section provides a hands-on demonstration of how the framework operates in identifying, analysing, and attributing Information Influence Operations.

Finally, the report concludes with a concise critical discussion that reflects on a number of trends regarding attribution, and the utility of the IIO attribution framework.

Published concurrently is a companion piece authored by researchers at Cardiff University, which provides guidance for open-source investigators and researchers involved in the process of attributing information influence operations.[1] It offers an account of generalizable principal issues that need to be thought about and factored in, when making decisions that impact upon attribution work using open-source data. Despite the differing emphases between the reports, they can be seen as complementary and share common ground by a joint perspective on attribution and exploring its application as related to information influence operations.

[1] Innes, M. & Ahonen, A. (2025). Attribution and Information Influence Operations: A 'Field Guide' for Open-Source Investigators and Researchers. ADAC.io EU Horizon Project Deliverable 1.2. Security, Crime and Intelligence Innovation Institute; Cardiff University.

2 The IIO Attribution Framework

Attribution refers to identifying the cause or origin of something. It involves making a determination about where something comes from or who is responsible for it. For example, specialty auction houses depend on accurate attribution to distinguish genuine high-value and exclusive items from forgeries. In the context of influence operations, attribution focuses on uncovering the actor behind the operation. This process is often complicated by factors such as limited access to critical information, intentional efforts to obscure evidence, and the challenge of separating the original creator of a campaign from those who may merely amplify or replicate its content. The task is further complicated by uncertainties about what constitutes acceptable deception in public discourse. In many liberal democracies, for instance, the rights and responsibilities of citizens or residents can differ significantly from those of foreign entities, such as international news organizations. The foreign dimension of foreign information manipulation and interference (FIMI) is often a decisive factor that prompts deeper investigation into suspicious material. However, identifying such a connection and crafting an appropriate response both depend on accurate attribution.[2]

In 2022 the NATO Stratcom CoE and Hybrid CoE IIO Attribution Framework (IIO Attribution Framework) was introduced, authored by Pamment & Smith.[3] A forthcoming report also aims to demonstrate how the framework can be applied to the Russian information manipulation and interference targeting Ukraine, including efforts to influence perceptions of the full-scale invasion of Ukraine in neighbouring countries.[4] They argue that attribution discussions concerning information influence operations (IIO) are heavily informed by principles rooted in cybersecurity, where technical evidence plays a central role in enhancing confidence in an attribution. Cybersecurity literature has long emphasized the primacy of technical analysis in identifying and understanding cyberattacks. Within this framework, hacker groups, often referred to as "Advanced Persistent Threats" (APTs), display distinct behavioural patterns that facilitate identification. The methodologies for assessing APTs are well-established and standardized, enabling the effective exchange of information regarding network vulnerabilities and attack vectors. Notable frameworks such as the Q Model,

---

[2] Ördén, H., & Pamment, J. (2021). What is So Foreign About Foreign Influence Operations? *Carnegie Endowment for International Peace*.

3 Pamment, J. & Smith, V. (2022). *Attributing Information Influence Operations: Identifying those Responsible for Malicious Behaviour Online*. NATO Stratcom CoE and Hybrid CoE.

4 Pamment, J., Smith, V. & Tsurtsumia, D. (forthcoming 2025*). Attributing Russian Information Influence Operations*. Ukraine Centre for Strategic Communications and Information Security and NATO StratCom CoE.

the ODNI Cyber Threat Framework, and The Diamond Model of Intrusion Analysis have been instrumental in refining best practices for cyberattack attribution.

2.1 The relationship between IIO and cyber attribution

Attribution of IIOs differs fundamentally from the technical focus typical of attribution in the cyber security realm, requiring a distinct methodology rooted in communication analysis. While cyber attribution often relies on clear technical markers—akin to identifying an artist by unique brushstrokes or layering techniques—attributing IIO demands a deeper examination of behavioural and contextual evidence. This includes assessing the content of messages, user accounts, narrative patterns, target audiences, communication strategies, and coordination efforts. These elements provide critical insights into the strategic intent behind such operations. Unlike cyberattacks, which usually involve clear illegal activities such as network intrusions and leave distinct technical traces, IIO operates within the public sphere, where legal boundaries are more ambiguous, and the roles of narrative originators and amplifiers often overlap.

The challenge of attributing IIOs is compounded by the generic nature of many observable behaviours and the difficulty of linking these behaviours to specific sources with certainty. Patterns of behaviour and discourse become the primary focus, but the evidence often lacks the specificity needed to conclusively identify an originator. Content analysis in an IIO reveals its tendency to polarize political opinions, disrupt public discourse, and influence decision-making, making its contextual impact central to the attribution process. These complexities highlight the nuanced and less definitive nature of IIO attribution compared to the relatively straightforward technical evidence often available in cybersecurity investigations.

Attribution of information operations (IIO) can be strengthened by access to secret intelligence, such as signal intercepts (SIGINT) or human intelligence (HUMINT), as well as proprietary data from digital platform backends—often referred to as telemetry. Insights derived from linking IP addresses, phone numbers, and email accounts to individuals or organizations further enhance the strength of investigations. However, in instances where such high-level resources are not available—either because the case does not meet the threshold for law enforcement or intelligence agency involvement—investigators are often left working with incomplete datasets. Despite these gaps, it is sometimes possible to establish links to specific countries, though identifying the responsible individuals or entities within those countries may remain elusive.

2.2 Categories of evidence

Attribution relies on three main categories of evidence: technical, behavioural, and contextual, all of which are considered alongside a legal and ethical framework. **Technical evidence** encompasses digital trails like IP addresses or other digital

markers left behind during operations. **Behavioural evidence** focuses on the methods and strategies employed, including Tactics, Techniques, and Procedures (TTPs). **Contextual evidence** examines the narratives and political dynamics surrounding the operation, such as the messages being spread and their intended impact. The legal and ethical framework ensures that these findings are approached with care, balancing considerations like proportionality, privacy, and broader geopolitical ramifications.

These evidence categories are further distinguished by the types of data sources used. Open source data, including publicly available research, APIs, and Open Source Intelligence (OSINT), is the most accessible. Proprietary data, such as intelligence derived from social media platforms or private sector analyses, provides additional layers of detail. Classified intelligence, such as SIGINT and HUMINT, is the most restricted and typically available only to select actors. Because access to these different types of data is so uneven, most public attributions rely heavily on open source and proprietary information. The limited availability of classified intelligence underscores the challenges of creating robust and definitive attributions in the public domain, particularly when addressing complex and multilayered influence operations.

2.3 Access to data

An attribution is shaped by the types of evidence available, which are typically drawn from three primary sources: open source, proprietary data, and classified intelligence. These sources not only influence what information is collected and how it is analysed but also determine the extent to which it can be made public.

Open source, accessible to any actor, relies on publicly visible content and is commonly used by NGOs, media, researchers, and intelligence agencies. Investigations using open source data often build circumstantial cases, combining techniques like qualitative and quantitative analysis of narratives and tactics with technical links, such as domain ownership or IP addresses. While this can provide valuable insights into an adversary's responsibility for an IIO, the technical data needed for robust attribution is rarely accessible through open sources. Moreover, the ethical obligation to inform the public often outweighs concerns about potential political or commercial repercussions.

Proprietary source, by contrast, is derived from privileged backend sources managed by digital platforms, private intelligence firms, data brokers, or cybersecurity companies. This data is particularly rich in technical and behavioural insights, often revealing the infrastructure and coordination behind an IIO. Attribution in this context typically takes the form of platform takedowns or intelligence reports, where actors and their activities are identified using proprietary and sometimes open source information. The legal assessments tied to such proprietary data are often shaped by platform terms of service as well as national laws, but decisions to attribute can also be influenced by commercial and geopolitical factors, such as market access or risks of retaliatory regulation. Despite the depth of information proprietary sources can provide, their

scope is limited to the data held by specific entities, often requiring partnerships for cross-platform analysis.

Classified source represents the most restricted origin of critical data, primarily accessed by governments and military entities. While it can incorporate open and proprietary data, classified intelligence is typically tailored to specific requests about adversarial activities. These assessments draw heavily on technical evidence but are often combined with behavioural, contextual, and legal analyses to provide a broader understanding of an actor's hostile actions. Such attributions can have significant consequences, including public denunciations, diplomatic actions, or retaliatory measures. However, the dissemination of classified intelligence is usually limited to internal government use or sanitized summaries shared with select audiences, and only occasionally released publicly in declassified formats.

The perspective and effectiveness of an analyst investigating IIOs are inevitably shaped by their access to these sources and their specific objectives, priorities, and resources. No single actor is likely to have unrestricted access to all three sources, nor the capacity to fully leverage them. Acknowledging the strengths, limitations, and potential gaps in each type of data is therefore critical for identifying biases, assessing probabilities, and ensuring a balanced and credible analysis.

2.4 The framework

An IIO attribution framework serves several purposes. First, it seeks to enhance mutual understanding among various actors—such as journalists, researchers, NGOs, companies, intergovernmental organizations, and governments—about the strengths and limitations of the information each has access to. Second, this improved understanding aims to foster better information sharing within the IIO community, thereby encouraging more effective collaboration.

Ultimately, these efforts contribute to the broader goal of equipping the public with the knowledge necessary to comprehend the nature of threats posed by IIOs. While structural challenges surrounding IIO attribution persist, the proposed framework does not claim to resolve all issues. Instead, it aims to highlight the core challenges, increase transparency, and suggest incremental yet feasible steps toward improvement.

The framework also emphasizes that a credible attribution is composed of multiple elements, which together provide a comprehensive picture. In some cases, certain categories of evidence may carry significant weight, while others may have little or no supporting data. Similar technical evidence might be gathered from both open and classified sources, but legal and ethical considerations—such as the need to protect sources—can sometimes prevent attribution, particularly when classified intelligence plays a key role.

Additionally, the framework's matrix offers a structured way to communicate and share generalized, high-level data about the factors involved in attribution. For instance, an actor could use the matrix to indicate that their decision was primarily based on open source contextual data, adding nuance to the attribution process and helping stakeholders understand the basis of the decision.

As illustrated by the examples in Chapter 4, the framework can enhance standard practices regarding measures against IIOs. It also has the potential to improve information sharing across the field, making these practices more effective and collaborative.

The matrix below presents the three kinds of evidence, as well as the supporting legal & ethical assessment, that are acquired from the three main data sources:

	Technical evidence	Behavioural evidence	Contextual evidence	Legal & ethical assessment
Open source	Web domain ownership, IP addresses, economic ties	Account activity, page activity, posting/cross-posting, sharing, follows, network	Media content, discourse and narratives, linguistics, political context, cui bono	Risk of litigation; research ethics; personal risk of becoming a target
Proprietary source	Data collected by platform backend	As above, with more extensive platform data	As above and data on previous takedowns with suspected links	Protecting political and commercial interests; data protection
Classified source	SIGINT; proprietary source data acquired by warrant	As above and SIGINT, HUMINT	As above and classified geo-political assessments	Actor-specific strategy; protecting political interests; data protection

Figure 1: IIO Attribution Framework

In the attribution analysis process, the evidence categories can be represented using traffic light signal colours to signify the strength of applicable indicators present. Green signifies strong indicators, yellow denotes moderate indicators, red represents weak indicators, and the absence of colour indicates no indicators within the category.

2.5 Technical evidence

Open source technical evidence is often more effective for analysing individual activities like disinformation than for investigating coordinated efforts such as IIOs. Analysts can use open source intelligence (OSINT) techniques, which include a range of readily available tools and methods, to gather technical and behavioural evidence. However, investigating coordination using only open sources is far more difficult compared to having access to internal infrastructure. For instance, an academic study on trolling in online news platforms revealed that the public API of such platforms only provided user location data based on IP addresses.[5] Since VPNs easily can mask a user's location, this limited data can lead to incorrect attributions and misunderstandings about adversarial methods. Researchers must often weigh the value of partial evidence against the risks of misinterpretation when better data is unavailable.

Open source evidence can help establish links between a social media account or webpage and its owner, uncovering hidden relationships between accounts and organizations. This information can be enriched by proprietary or leaked classified data, such as mobile phone records, passport applications, or airline bookings, which have been used in investigations by organizations like Bellingcat.[6] When such evidence is made public, it greatly strengthens the foundation for attributing actions to specific actors. However, the use of leaked datasets is more common among investigative journalists than academic researchers.

Proprietary technical evidence is rarely made public and typically relies on data collected and analysed by platform owners. Yet, these companies face internal challenges in managing and interpreting this data, which are often not well understood outside the industry. Such challenges include consolidating information from disparate sources, assigning responsibilities for analysing specific behaviours across teams or departments, and ensuring sufficient staffing to handle these tasks effectively.

For example, data about the creation of new accounts can reveal details about the account creator, their location, and potentially other online activities. Cross-referencing this data may uncover patterns that indicate a coordinated effort to establish an infrastructure capable of executing an IIO. Platforms also deploy countermeasures and traps to combat persistent threat actors in an ongoing effort to deter malicious activity.

When analysis is shared publicly, it is usually presented as a conclusion, such as attributing certain activities to individuals linked to military operations in specific countries. A Stanford report on GRU influence operations, for instance, attributed

5 Zelenkauskaite, A., & Balduccini, M. (2017). “Information Warfare” and Online News Commenting: Analyzing Forces of Social Influence Through Location-Based Commenting User Typology. *Social Media + Society*, 3(3).

6 Bellingcat. (2020). *Navalny FSB Methodology*. Retrieved from https://www.bellingcat.com/resources/2020/12/14/navalny-fsb-methodology/

findings to Facebook's analysis,[7] while the Graphika report *From Russia with Blogs* noted that it could not independently verify a social media platform's attribution of actions to Russian military intelligence due to insufficient evidence.[8] Data of this kind is often shared with law enforcement or intelligence agencies when legal frameworks permit. Additionally, private intelligence firms, which are increasingly active in the IIO domain, sometimes share proprietary reports with journalists to gain market credibility or pursue political objectives.

Classified technical evidence is sourced from intelligence agencies through methods like signals intelligence (SIGINT) and human intelligence (HUMINT). These covert methods allow agencies to collect detailed evidence that can link adversaries to specific IIO. Similar to proprietary evidence, classified findings released to the public are often summarized as conclusions to protect personal data and the methods of collection. For instance, in the Mueller Report on Russian interference in the 2016 US Presidential election, 40% of the 448 pages contained redactions. The sections on Russian IIO were the most heavily redacted, with 46% of the content withheld—more than in sections discussing prosecution decisions or hacking activities.[9]

Financial data represents an important sub-category of technical evidence in the attribution of malign influence operations. They complement key indicators such as contextual, behavioural, and other forms of technical data by providing unique insights into the financial activities and ties underlying such operations. Analysing financial data and trails can help differentiate between various threat actors and establish connections that point to state authorities or other controlling entities. The integration of financial signals into the broader technical evidence framework enhances the ability to attribute actions or events with greater confidence, particularly when these signals align with content, behavioural, and other technical data.

The ability to confidently attribute actions or events is enhanced when multiple categories of evidence converge to point to a specific source. However, the practical challenge of attribution remains significant, as access to all types of evidence—content is rarely available. In many cases, assessments must rely on only one or two forms of data, which inherently limits the accuracy and precision of the conclusions drawn. This limitation underscores the importance of incorporating financial signals into a comprehensive analytical approach, where their role within technical evidence can provide a critical layer of depth and reliability in identifying the actors behind foreign information manipulation and interference.[10]

---

7 DiResta, R., & Grossman, S. (2020). *Potemkin Think Tanks.* Stanford Internet Observatory, Freeman Spogli Institute for International Studies, Stanford University.

8 Nimmo, B., François, C., Eib, C. S., & Tamora, L. (2020). *From Russia With Blogs.* Graphika.

9 Zarracina, J., & Chang, A. (2019). Mueller Report Redactions, in One Chart. *Vox*.

10 Innes, M. & Ahonen, A. (2025). Attribution and Information Influence Operations: A 'Field Guide' for Open Source Investigators. ADAC.io EU Horizon Project Deliverable 1.2. Security, Crime and Intelligence Innovation Institute; Cardiff University.

2.6 Behavioural evidence

Behavioural evidence refers to patterns and indicators in actions that may point to inauthentic behaviour, attempts to obscure a campaign, or evidence of coordination. This can include practices such as trolling, creating sensationalist content to attract attention, presenting fake experts, and generating content designed to divide audiences. Collecting this type of evidence involves closely examining the ways communication strategies, influence methods, and specific actions combine to reveal the workings of a campaign. This is often categorized under Tactics, Techniques, and Procedures (TTPs), which provide a structured way to understand threat actor behaviours.

According to The U.S. National Institute of Standards and Technology (NIST), a leading body on cybersecurity standards, TTPs refer to the behaviour of a threat actor as it is observed through their actions. A tactic represents the overarching goal or objective, such as spreading false information about an event. Techniques detail the specific methods used to support this goal, for instance, presenting fabricated video footage to deceive an audience reliant on television news. Procedures offer a granular breakdown of how these techniques are executed, such as the technical steps involved in creating deceptive visuals, manipulating settings, and circulating the content through social platforms before its adoption by established media-channels.

Open source behavioural evidence often concentrates on the activities and methods of accounts suspected to be part of an IIO. Analysts typically examine factors such as posting times, amplification methods like coordinated sharing, and connections between accounts identified through network behaviour like liking, following, or sharing. Many of these insights can be obtained using widely available analytical tools designed for open source intelligence.

Proprietary behavioural evidence enhances this type of analysis by drawing on data accessible only to platform owners, such as private account activity or content from closed groups. This approach often reveals broader patterns, including how actors circumvent platform security measures. Similarly, **classified behavioural evidence** uses comparable methods but includes intelligence from additional sources, such as intercepted communications or financial transactions, to uncover relationships between suspected actors and illicit activities.

An analysis of the methods used in reports studied indicates that behavioural evidence is a cornerstone of attribution. Frequently employed techniques include assessing social media accounts to understand their creation details, profile characteristics, and posting content; analysing interactions with other accounts to map networks; and examining the timing of account activity. These approaches allow researchers to identify recurring patterns, which are crucial for connecting specific actions to their sources. The prominence of behavioural evidence is underscored by the fact that five of the seven most frequently applied methods focus on this area.

Behavioural analysis extends to examining the communication strategies of adversaries. Investigating the types of messages they craft and the methods they use to disseminate them helps analysts understand their objectives and the mechanisms behind their actions. This deeper understanding of adversary communication patterns provides a clearer view of their intentions and tactics.

2.7 Contextual evidence

This category of evidence focuses primarily on the content circulated during influence operations, often involving narratives. Narratives are essentially simplified stories that help people make sense of complex issues, shape perceptions, and convey ideas about identity, community, and purpose. They typically represent collective beliefs rather than literal truths, reflecting values and ideas developed over time within a community. Understanding these narratives includes examining the values and identities they express, the audiences they resonate with, and the sources of their credibility.

Narratives are especially significant when connected to specific communities or real-world events. Disinformation can play a central role in building adversarial narratives. For instance, the pro-Kremlin narrative that Ukraine poses an existential threat to Russia may be reinforced by false claims, such as allegations of secret labs linked to Covid-19. Contextual evidence encompasses not just the overarching story but also details such as false statements that align with and reinforce broader narratives, the strategic timing of messages in relation to significant events like elections or major political decisions, and the connection of contemporary messaging to historical manipulation techniques. They also frequently leverage concerns specific to the targeted communities, aiming to deepen their resonance. Analysts examine who benefits from these narratives and who is harmed, identify the intended audience, and assess how closely these narratives align with publicly stated positions of governments or officials.

This type of evidence also aids in analysing the motivations and goals behind influence operations (IIOs), such as electoral interference or societal polarization. However, it remains one of the most challenging areas to study due to its subjective and culturally dependent nature. IIOs often exploit cognitive biases, and their provocative content can be difficult to categorize. By integrating contextual evidence with behavioural insights, analysts can evaluate who benefits ("cui bono") and forecast malign intentions. Nevertheless, this approach can be misleading, as actors sometimes craft operations to falsely attribute origins to others.

Unlike forensic evidence in cyberattacks, which can confirm whether an intrusion occurred, influence operations can be subtle and prolonged. Content may be seeded across various platforms over time or emerge spontaneously. Both approaches often blend seamlessly with genuine issues. For example, in 2018, Facebook dismantled a group created by an adversary to organize a real public demonstration, illustrating how

influence operations exploit existing grievances rather than directly causing events. The impact of such operations lies in their ability to distort narratives, introduce disinformation, and undermine trust in public discourse.

Open source contextual evidence focuses on analysing digital content—text, images, videos, hashtags, and language—within a geopolitical framework to uncover an IIO's intentions and effects. Questions may include whether the operation seeks to polarize debates, discredit actors, suppress voting, or affect territorial defence. Analysts rely on platform data from public Application Programming Interfaces (APIs), such as Meta's now discontinued CrowdTangle, and use these alongside technical and behavioural evidence to establish a case linking the operation to specific actors and goals.

Proprietary and classified data sources provide further insights, including access to metadata and hidden accounts. Investigative journalists or analysts with access to proprietary tools may infiltrate closed groups or leverage information from intelligence agencies. These sources offer broader datasets, strengthening contextual analysis and attribution efforts.

Building a robust evidential case requires a thorough understanding of technical and behavioural elements of IIOs. Analysts map networks across platforms, identifying key accounts driving narratives and those amplifying them. These networks may involve both coordinated actors and unsuspecting individuals engaging with the content. Attribution often remains ambiguous, as seen with entities like the Internet Research Agency (IRA). While linked to Russian interests, the IRA's ownership and relationship with the Russian government were deliberately obscured, enabling plausible deniability for state actors. Attribution language aims to describe these connections without overstepping available evidence while safeguarding investigative methods.

Contextual analysis also considers the linguistic and cultural elements of influence operations. Analysts examine the languages and dialects used, whether hashtags are borrowed or created, and how narratives align with local grievances or events. They investigate whether content connects to offline activities, such as businesses or events, and explore questions about beneficiaries and targets. Effective analysis involves identifying information gaps and employing appropriate tools to address them, ensuring a comprehensive understanding of the influence operation's scope and impact.

2.8 Legal and ethical assessment

Legal and ethical assessment is less a part of the analytical process and more a supporting assessment and continuous set of considerations analysts must address while gathering evidence. For instance, is it ethical for an analyst to join a private social media group using a pseudonym? The answer depends on whether the analyst is a journalist, academic researcher, or civil servant, as different ethical standards and legal frameworks apply. Actions acceptable under one jurisdiction or role may be legally or

politically problematic in another, especially if evidence is intended for litigation. For example, an analyst might choose not to publicly name an EU-based individual spreading pro-Kremlin messages due to freedom of expression protections. However, direct links to a Russian company might justify public attribution, even amid such concerns. Governments may hesitate to make attributions public to protect sensitive sources, while civil society must weigh the benefits of disclosure against potential retaliation by threat actors, making these decisions inherently complex.

These considerations influence the language used in attributions to minimize legal and reputational risks. The choice of wording also depends on the evidence's nature and source. Open source investigations raise ethical questions about proportionality and methodological soundness, especially regarding leaked data. Legal risks, including libel, vary across jurisdictions, and researchers may face platform bans if their methods breach platform rules, which platforms enforce with inherent conflicts of interest.[11] Individual journalists and researchers often face personal safety risks when conducting investigations.

Privacy is a key concern for organizations handling proprietary evidence. Digital platforms are wary of exposing identifiable personal data, particularly under regulations like GDPR. While GDPR includes research exemptions, platforms often cite privacy concerns to withhold data, sometimes influenced by geopolitical or commercial interests, such as regulatory risks, market access, and advertising revenues.[12] For instance, Facebook's focus on IIO in Western regions and adversarial states like Russia and Iran reflects such strategic priorities, as a 2021 whistleblower case appears to confirm.[13] It also plays a crucial role in handling classified data. Intelligence agencies investigate actors to assess their behaviour and intentions, sharing attributions among allies to align strategies and address problematic international actions. Public attributions may serve as part of actor-specific strategies or threat intelligence statements, both to inform and deter. Governments often make attributions public to justify legal actions, such as sanctions or expulsions, emphasizing their importance in diplomatic and security contexts.

In the realm of attribution, the processes and standards vary significantly between social media platforms versus the field of journalism and the civil society sector, highlighting contrasting approaches to transparency and accountability. Social media platforms often attribute actions when removing networks of accounts that breach their

[11] Erwin, M. (2021). Why Facebook's Claims About the Ad Observer Are Wrong. *Mozilla Blog*. Retrieved from https://blog.mozilla.org/en/mozilla/news/why-facebooks-claims-about-the-ad-observer-are-wrong/

[12] Gillum, J., & Elliott, J. (2021). Sheryl Sandberg and Top Facebook Execs Silenced an Enemy of Turkey to Prevent a Hit to the Company's Business. *ProPublica*.

[13] Wong, J. C. (2021). How Facebook Let Fake Engagement Distort Global Politics: A Whistleblower's Account. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2021/apr/12/facebook-fake-engagement-whistleblower-sophie-zhang

policies. However, these platforms rarely disclose the technical evidence underpinning their decisions, leaving their methodologies opaque. Furthermore, their investigations are confined to activity on their own platforms, overlooking cross-platform behaviours, which creates gaps in the broader understanding of influence operations and negatively affects the ability of civil society actors to assess their methodologies.[14]

In contrast, journalism and non-governmental organizations have developed systematic frameworks to evaluate source credibility and ensure transparency. Initiatives like NewsGuard assess the trustworthiness of news websites based on defined criteria,[15] while the ICFN Code of Principles[16] guides fact-checking organizations in producing impartial and rigorous evaluations of public claims. Tools like DFRLab's Foreign Interference Attribution Tracker,[17] an interactive, open source database documenting allegations of foreign interference in U.S. elections, exemplify such efforts to evaluate the evidence and overall impact of such attributions in a transparent manner.

Public attributions are also influenced by the perceived severity of influence operations. For example, in March 2018, the poisoning of Sergei and Yulia Skripal in the UK with the nerve agent Novichok, which later harmed two additional individuals, prompted the UK government to release a detailed investigation attributing the attack to named Russian GRU officers. This incident highlights how governments may disclose detailed evidence when the public threat is significant, as in this extreme case of state-sponsored violence.

2.9 Confidence intervals

Determining attribution is inherently complex, meaning assessments often lack absolute certainty and rely instead on probabilities. Analysts typically use concepts like the balance of probability to express their conclusions. Confidence intervals, commonly employed in intelligence analysis, reflect the inherent uncertainty and likelihood of assessments rather than presenting them as definitive facts. For instance, analysts may review extensive sources, such as HUMINT and SIGINT, to conclude that the probability of a terrorist attack in a capital city is low. These conclusions are based on risk assessments rather than absolute certainty. Similarly, attributing influence

[14] Pamment, J., & Ahonen, A. (2024). *The Ethics of Outsourcing Information Conflict: Outlining the Responsibilities of Government Funders to Their Civil Society Partners.* NATO Strategic Communications Centre of Excellence.

[15] NewsGuard. (2024). *NewsGuard Ratings: Rating Process & Criteria.* Retrieved from https://www.newsguardtech.com/ratings/rating-process-criteria/

[16] International Fact-Checking Network (IFCN). (2024**).** *ICFN Code of Principles: The Commitments.* Retrieved from https://www.ifcncodeofprinciples.poynter.org/the-commitments

[17] DFRLab. (2024). *Foreign Interference Attribution Tracker.* Digital Forensic Research Lab, Atlantic Council. Retrieved from https://interference2024.org/

operations involves analysing numerous data points that together form a probabilistic picture rather than a definitive judgment.

Probabilities can be communicated through percentages or descriptive terms, but these can sometimes lead to misinterpretation. The UK Government's "Probability Yardstick"[18] combines percentages with descriptors, such as "unlikely" for 25-35% or "highly likely" for 80-90%. Similarly, the Oasis Open project which governs the use of STIX confidence objects[19] offers options like "low, medium, high," numerical ranges, or descriptive scales such as the Admiralty Credibility Scale, DNI Scale and Words of Estimative Probability. Oasis Open allows these scales to be used interchangeably by aligning numerical ranges with descriptors, ensuring flexibility across different analytical frameworks. Clear definitions of each category are crucial to minimize errors and ensure consistent interpretation.

More critical than the specific scale chosen is the need for precise definitions for each category to reduce human error and ensure clear, reliable interpretation. This clarity is essential to maintain consistency and avoid misunderstandings when interpreting probabilities.

Attributions in influence operations are particularly challenging because varying levels of confidence and attribution types may apply to accounts within the same network. For example, analysts may confidently attribute a statement from an official government social media account to a nation-state. However, it may be far harder to attribute the same confidence level to another account in the network that poses as an individual sharing government content.

Consequently, when analysing and presenting findings on IIOs, it is important to acknowledge that not every account in a network can be attributed with certainty. For those accounts where attribution is possible, confidence levels may vary considerably. Clearly articulating the degree of confidence and the extent of attribution can help clarify what is established, what remains uncertain, and avoid assumptions unsupported by evidence.

[18] UK Ministry of Defence. (2023). *Defence Intelligence – Communicating Probability.* Retrieved from https://www.gov.uk/government/news/defence-intelligence-communicating-probability

[19] OASIS Open Standard. (2021). *STIX Version 2.1, Appendix A.* Retrieved from https://docs.oasis-open.org/cti/stix/v2.1/os/stix-v2.1-os.html#_1v6elyto0uqg

3 Overview of methodologies and organizations related to attribution

Attribution methodologies serve as critical tools in the broader effort to counter foreign information manipulation and interference (FIMI). In an era defined by the pervasive influence of digital ecosystems, the ability to identify and contextualize actors engaged in malign activities is not merely a technical challenge but involves developing structured approaches combining technical tools and human-driven analysis and assessment. This chapter will revisit and provide an update to the literature review conducted in conjunction with the creation of the IIO attribution framework and adds an overview of methodologies utilized by organizations engaged in this domain.

Worth bearing in mind is that this overview is limited by the fact that government and alliance entities involved in analysis and decision-making relating to attribution of information influence operations rarely publicise their methodologies at any depth. This is also true for certain threat intelligence and cybersecurity companies. Understandable, of course, due to the need to not expose respective organizations' tactics, techniques and procedures for strategic or competitive reasons, but something which of course to some extent limits getting a complete picture of the whole range of methodologies used.

Microsoft and Recorded Future, two of the leading entities engaged in the threat intelligence space, have incorporated structured attribution frameworks into their processes to enhance the identification and analysis of FIMI activities. Both organizations have adapted the IIO framework presented in this report and applied them to their workflows to suit their capacities and objectives.[20][21] Hence, they will be explored in greater depth in Chapter 4 that covers how the IIO attribution framework has been used to date.

[20] Microsoft Threat Analysis Center. (2023). *An attribution model for influence operations*, p. 2. MTAC White Paper.

[21] Recorded Future, Insikt Group. (2024). *"Operation Undercut" Shows Multifaceted Nature of SDA's Influence Operations,* p. 7. Recorded Future.

3.1 Literature review findings

The 2022 publication of the IIO Attribution Framework was preceded by an analysis of 59 reports on information influence operations (IIO)[22] authored by 24 different organizations and data from the Disinfodex database.[23] A significant proportion of the reports, nearly half, were produced by three prominent U.S.-based organizations: Graphika, The Atlantic Council and its Digital Forensic Research Lab (DFRLab), and Stanford's Internet Observatory. This highlights not only their extensive contributions to the field but also the dominance of American research institutions in this area. These organizations benefit from unique data-sharing arrangements with platforms like Facebook and Twitter, granting them advance notice of takedown actions. This advantage enables the synchronization of their report publications with platform takedown announcements. However, in such cases, their research often relies on platform-provided attributions, which they attempt, though not always successfully, to corroborate independently.

The study found that 19 of the reports included direct attributions, 26 referenced attributions made by external sources, and 10 augmented external attributions with their own evidence. Only four reports did not feature any attribution. Among the actors identified, Russian entities accounted for the largest share (49%) of attributions in the reports analysed, with Iranian actors coming next at 12%.

The reliance on proprietary data from platforms, particularly Twitter (66%) and Facebook (59%), emerged as a central theme in the reports, serving as the primary source of technical and behavioural information for attribution. However, the technical data underpinning these attributions is not shared with independent researchers, leaving most attributions reliant on interpretations of proprietary data through the lens of publicly available information. This raises methodological concerns about the reliability of recreating platform takedowns based solely on open sources.

Governments, like platforms, are generally reticent to disclose detailed assessments of IIO. A notable exception was the publication of evidence related to Russian interference in the 2016 U.S. Presidential election, as part of an FBI inquiry. More

[22] Pamment & Smith reviewed 88 reports on IIO by think tanks, universities, governments, companies, and other research organizations from 2017–2020. Reports focusing on general trends in IIO or those not detailing specific IOs were excluded from the sample. This resulted in 59 reports being selected for further analysis. The content of each report was analysed and coded to capture information including basic metadata about the report, the source of the data, research techniques employed, tactics identified as used by an IIO, whether an attribution was made or inherited (e.g., from a platform takedown or other research), details of the suspected actor, languages used, specific geographic regions or countries targeted, and any assessments of the objectives or motives of the IIO. In total, 14 research techniques and 41 distinct tactics were identified.

[23] Disinfodex is a database which records Facebook, Google, YouTube, Reddit and Twitter takedowns. It complements the data from the independent reports by more comprehensively representing platform takedown actions, https://disinfodex.org

commonly, governments issue broad statements on the intent of hostile actors without providing substantive details. In some cases, heavily redacted reports create a misleading impression that analysts were working with limited data.

The Disinfodex online database documents 520 platform takedowns across Facebook, Google, YouTube, Reddit, and Twitter.[24] Approximately two-thirds of these takedowns include explicit attributions to actors. However, attributions concerning sensitive entities, such as governments or political parties, are often couched in ambiguous terms like "individuals associated with" or "employees of." This language reflects the influence of political and ethical considerations in attribution framing but often lacks clarity. Terminological inconsistencies, such as the use of vague phrases like "Kremlin-backed," obscure the extent and nature of government involvement. Efforts to address this issue, such as Jason Healey's "Spectrum of State Responsibility",[25] have sought to establish a more precise framework for categorizing state roles in IIO, yet challenges remain in articulating these nuances effectively.

Building on the initial analysis, an updated review covering reports from 2022 to 2025 expands the scope to include 38 additional reports authored by a diverse array of organizations, including Recorded Future, Microsoft Threat Intelligence, EUvsDisinfo, Debunk.org, Mandiant, and the Institute for Strategic Dialogue (ISD).[26] The review found that direct attributions were made in 19 of the reports, referenced attributions appeared in 3, and 11 reports augmented external attributions with their own evidence. In 3 reports, no attribution was made. Among the actors identified, Russian entities continued to dominate, accounting for 64 % of attributions, followed by Chinese state-linked actors at 18% and Iranian actors at 12%. Other entities, including Venezuela, Belarus, and private firms linked to influence campaigns, were also featured. Platforms such as Telegram (44%), Facebook (38%), X (former Twitter) (32%), YouTube (18%), and TikTok (12%). The creating of fake websites posing as legitimate news outlets was also notable (17%).

As highlighted in the initial analysis, attributions involving sensitive entities, such as governments or political parties, are often framed in vague terms such as "state-linked" or "affiliated." For instance, one report referenced "Kremlin-backed narratives" and

[24] Disinfodex. (n.d.). *Disinfodex Database of Platform Takedowns.* Retrieved from https://disinfodex.org

[25] Healey, J. (2012). *Beyond Attribution: Seeking National Responsibility for Cyber Attacks.* Atlantic Council. Retrieved from https://www.atlanticcouncil.org/wp-content/uploads/2012/02/022212_ACUS_NatlResponsibilityCyber.PDF

[26] The updated review for this report analysed 47 publications on Information Influence Operations (IIO) released between 2022 and 2025 by think tanks, universities, companies, and other research organizations. Following the methodology established by Smith and Pamment, the reports that lacked detailed information about specific operations or focusing solely on general IIO trends were excluded which resulted in a final sample of 38 reports. These reports were systematically coded to capture essential details, including the methods used, data sources, whether attribution was made, and, if so, the type of attribution (direct, referenced, or augmented with additional evidence), as well as information about suspected actors and the motives for the IIO.

another refrains from making a definitive attribution but still refers to the activity as "alleged Russian involvement".

The reliance on open source data was evident, with many reports combining multiple types of data, including technical evidence, behavioural evidence, and contextual evidence. Some attributions were also made by combining open source data with proprietary evidence like private emails or access to internal platform data. This differs from the previous analysis where reliance on proprietary data from platforms emerged as a central theme and served as the primary source of technical and behavioural information for attribution. It appears that obtaining proprietary data from platforms may now be more challenging. In the updated analysis, direct attributions made with a combination of open source and proprietary data often relied on insights based on access to internal data within the specific organizations. In cases of referenced and augmented external attributions where proprietary evidence played a central role, the attribution was frequently tied to platform takedown reports or attributions or analysis made by organizations that had access to proprietary data. As such, a lot of the technical data that underpin these attributions remains inaccessible to independent researchers and leaves many attributions to be dependent on other actors' interpretations of proprietary data.

3.2 East StratCom Task Force and EUvsDisinfo

The European Union began addressing disinformation as a key security priority after Russia's use of disinformation emerged as a central tactic of hybrid warfare during the annexation of Crimea in February 2014. Responding to concerns raised by a group of member states, the European Council in March 2015 underscored the importance of countering Russia's persistent disinformation efforts.[27]

This call to action prompted the creation of the East StratCom Task Force within the European External Action Service's Strategic Communications Division. The task force was designed to strengthen the EU's messaging in its eastern neighbourhood, enhance the regional media landscape, and promote media freedom while supporting independent journalism in both neighbouring states and within the EU.[28]

Later in 2015, the task force established the EUvsDisinfo initiative to systematically identify, monitor, and analyse instances of Russian disinformation. The aim is to foster a deeper understanding of the tactics, narratives, strategies and objectives behind such activities among the public to enable citizens in Europe and beyond to develop

[27] European Council. (2015). *European Council Meeting (19 and 20 March 2015) – Conclusions.* Retrieved from https://www.consilium.europa.eu/media/21888/european-council-conclusions-19-20-march-2015-en.pdf

[28] European Commission. (2018). *Action Plan Against Disinformation.* Retrieved from https://www.eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf

resistance to digital information and media manipulation. Through its database,[29] which as of late 2024 contains over 18,000 cases, the Task Force has created a valuable resource for analysing the narratives and methodologies underpinning Russian disinformation efforts.

The Task Force operates with a clear mandate to identify, document, and debunk disinformation linked to Russian state actors. A work guided by the EU's Action Plan against Disinformation, which defines disinformation as "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm".[30] The database is curated according to three strict criteria: a claim must have a direct or indirect connection to Russian state actors, be verifiably false, and be disseminated with malign intent. Subjective statements or factual inaccuracies without intent to cause harm are excluded. An important distinction is that the mandate of the East StratCom Task Force is limited to media outlets connected to Russian state actors. An important distinction is that the Task Forces data gathering and publication of cases is confined to media outlets connected to Russian state actors. Claims originating from EU individuals or media outlets are excluded unless they can be explicitly traced to entities owned or managed by Russian state actors or affiliates.

The database is manually compiled with input from a network of contributors spread across EU member states and partner nations. Each case is paired with a comprehensive debunk that outlines the lack of factual validity in the claim. This meticulous process not only challenges individual instances of disinformation but also creates a rich repository of data for in-depth analysis. Patterns of targeting are evident, with Ukraine being the most frequently targeted nation, accounting for approximately one-third of the cases. Within the EU, Germany and Poland are frequently targeted, and the Baltic states are disproportionately affected relative to their population sizes.[31]

Beyond documenting individual cases, the Task Force's work helps illuminate the overarching strategies employed in Russian disinformation campaigns. This activity forms part of a larger hybrid threat that leverages a range of tools, actors, and platforms to meet its objectives. By examining recurring themes and keywords, the database sheds light on key narratives and highlights the focus on particular regions or issues, contributing to a deeper understanding of the operational scope and intent behind these campaigns.

29 EUvsDisinfo. (n.d.). *Disinformation Cases Database.* Retrieved from https://euvsdisinfo.eu/disinformation-cases/ (Accessed: November 30, 2024)

30 European Commission. (2018). *Action Plan Against Disinformation.* Retrieved from https://www.eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf

31 Enerud, P. (2022). *Narrating Disinformation: The Templates for Kremlin Lies*. Stockholm Centre for Eastern European Studies.

3.3 Microsoft Threat Intelligence

Over the past decade, Microsoft states that their threat intelligence team has tracked more than 300 threat actors, including 160 nation-state groups and 50 ransomware operators, and shared its findings with customers. An international group of analysts, penetration testers, data scientists, and experts in geopolitics and disinformation supports these efforts, using a broad, adversary-focused approach. In 2023, Microsoft introduced a new taxonomy for naming threat actors, drawing on weather-related nomenclature, where five primary categories are employed:[32]

Nation-state actors: Cyber operators acting on behalf of, or under direction from, a nation or state-aligned program, regardless of the underlying objective (espionage, financial, retaliatory, or otherwise).

Financially motivated actors: Criminal groups or individuals orchestrating cyber campaigns driven by financial gain, who cannot be definitively linked to a known nation-state or commercial entity. This category covers ransomware operators, business email compromise activities, phishing groups, and other forms of financially or extortion-motivated threats.

Private sector offensive actors (PSOAs): Commercially registered entities that develop and sell cyberweapons to various clients, who then decide on specific targets. These targets often include dissidents, human rights defenders, journalists, and civil society advocates.

Influence operations: Information campaigns, disseminated either online or offline, designed to shape perceptions, behaviours, or decisions among specific audiences in a manipulative manner.

Groups in development: Newly detected or evolving threat activities that remain under investigation until further information allows clearer attribution.

[32] Lambert, J. (2023-04-18) *Microsoft shifts to a new threat actor naming taxonomy*. https://www.microsoft.com/en-us/security/blog/2023/04/18/microsoft-shifts-to-a-new-threat-actor-naming-taxonomy/

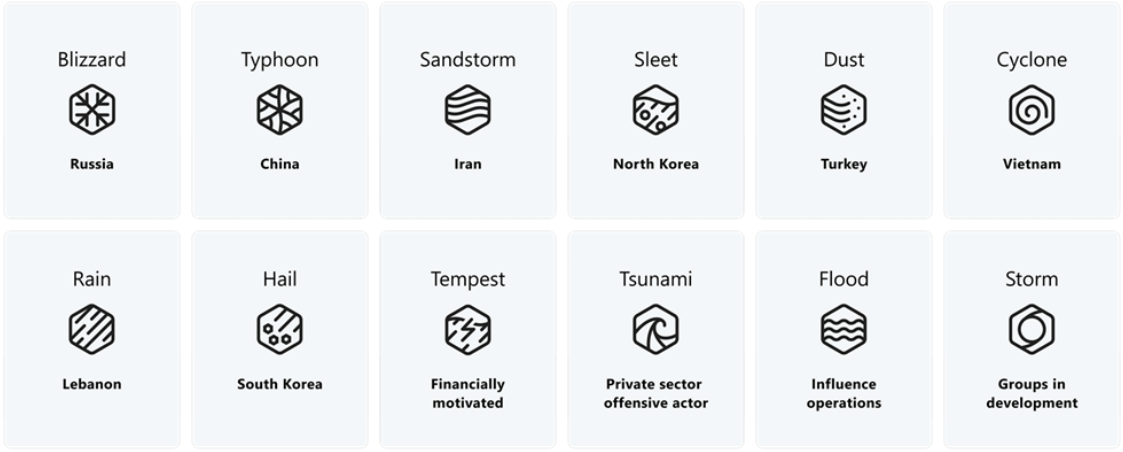


Figure 2: Microsoft threat actor naming taxonomy outlining different state actors and the additional four naming categories [33]

Microsoft employs an internal process to organize and track these "Groups in development." Until they can be merged with known actors or consolidated into newly named entities, they are referred to with the placeholder "Storm," a designation applicable to any actor type—nation-state, financially motivated, PSOA, or influence operation. The attribution process focuses on establishing each actor's infrastructure, tools, target selection, and motivation.

When a new threat emerges, Microsoft's Defender Threat Intelligence platform[34] supports the investigations by compiling available details into Intel Profiles.[35] These profiles classify threat actors by location, targeted industries, and methodologies, and include known Indicators of Compromise (IoCs)[36] and observed Tactics, Techniques, and Procedures (TTPs).

A recent illustration of this process is the analysis of threat actor Storm-0558 that utilized forged authentication tokens to gain access to user emails in circa 25 organizations—including government agencies—and related consumer accounts hosted on the public cloud. Microsoft's assessment landed in the threat actor originating in China.[37] Among the indicators guiding this conclusion were working hours consistent with those in China, as well as tactics resembling those used by other

[33] *How Microsoft names threat actors*. https://learn.microsoft.com/en-us/defender-xdr/microsoft-threat-actor-naming

[34] *Microsoft Defender Threat Intelligence*. https://www.microsoft.com/en-us/security/business/siem-and-xdr/microsoft-defender-threat-intelligence

[35] Mercer, D. (2023-03-29). *What's New: Intel Profiles Deliver Crucial Information, Context About Threats*. https://techcommunity.microsoft.com/blog/defenderthreatintelligence/whats-new-intel-profiles-deliver-crucial-information-context-about-threats/3780076

[36] *What are indicators of compromise (IOCs)?* https://www.microsoft.com/en-us/security/business/security-101/what-are-indicators-of-compromise-ioc

[37] Microsoft Threat Intelligence (2023-07-14). *Analysis of Storm-0558 techniques for unauthorized email access*. https://www.microsoft.com/en-us/security/blog/2023/07/14/analysis-of-storm-0558-techniques-for-unauthorized-email-access/

Chinese groups, including Violet Typhoon (also referred to as ZIRCONIUM or APT31). By examining the group's use of forged authentication tokens to access email accounts, Microsoft analysts were able to discern the threat actor's techniques and infrastructure, contributing to the overall attribution.

3.4 Google Threat Intelligence

In 2022, Google acquired the cybersecurity firm Mandiant, which now functions as a subsidiary of Google Cloud. This acquisition integrated Mandiant's team of security and intelligence professionals, who operate in 22 countries and serve clients across 80 countries.[38] Google Threat Intelligence provides a unified interface that consolidates information on threat actors, allowing users to refine their searches by source, targeted industry, region, motivation, and associated malware or tools.

The Google Threat Intelligence platform offers a list of associated objects, which can be filtered by modification date, object type (e.g., campaigns, malware, toolkits, collections, and vulnerabilities), origin (Google Threat Intelligence or partner sources), source regions, targeted regions, and targeted industries.[39] It also presents Indicators of Compromise (IoCs)[40] alongside telemetry data and detection rules based on the signatures of the cybersecurity tools Yara, Sigma, and IDS. Additionally, it includes a tab that outlines tactics, techniques, and procedures (TTPs) mapped to the MITRE ATT&CK framework.[41]

Google Threat Intelligence and Mandiant have defined three attribution categories for suspected threat actors: **Mandiant Confirmed**, **Mandiant Suspected**, and **Possible Association**.[42] Activity with sufficient evidence to conclusively link it to a known APT (state-sponsored), FIN (financially motivated), TEMP or UNC (uncategorized) group[43] is labelled **Mandiant Confirmed**. When available data suggests a potential relationship to an existing group but does not support full confidence, the cluster is designated as either Mandiant Suspected or Possible Association. **Mandiant Suspected** implies high or moderate confidence in the connection, whereas **Possible**

[38] Google Cloud (2022-09-12). *Google + Mandiant: Transforming Security Operations and Incident Response*. https://cloud.google.com/blog/products/identity-security/google-completes-acquisition-of-mandiant

[39] *Google Threat Intelligence Platform Navigation*. https://gtidocs.virustotal.com/docs/google-threat-intelligence-navigation

[40] Indicators of compromise serve as forensic evidence of potential intrusions on a host system or network.

[41] MITRE ATT&CK® is a knowledge base of adversary tactics and techniques. https://attack.mitre.org/

[42] *Google Threat Intelligence – suspected attribution*. https://gtidocs.virustotal.com/docs/suspected-attribution

[43] Vanderlee, K. (2020-12-17) Deb*UNCing Attribution: How Mandiant Tracks Uncategorized Threat Actors*. https://cloud.google.com/blog/topics/threat-intelligence/how-mandiant-tracks-uncategorized-threat-actors

Association indicates a low-confidence link. Both categories reflect the degree of overlap between the new activity and a known threat group. For instance, Mandiant Suspected may be applied when there are multiple overlapping data points of strong quality, while Possible Association may result from fewer similarities, parallels in TTPs, or other identifiable commonalities.

3.5 International Institute for Strategic Studies

The International Institute for Strategic Studies (IISS), a Defence and National Security think tank, has introduced the Degrees of Cyber Attribution framework to characterize how closely a cyber actor is connected to a government. The framework accounts for the varying nature of these ties and provides terminology to capture each relationship's nuances, recognizing that assessments can be refined over time as new details emerge.[44]

Non-state actors are those not clearly linked to any government through intelligence assessments. This group also includes entities affiliated with terrorist organizations, contractor or hacker-for-hire services, and certain cybercriminal networks.

State-linked actors are believed to have some relationship with a government, although the available evidence does not conclusively establish official involvement.

State-shaped actors carry out operations whose targets, interests, or capabilities mirror those of a government, yet analysts lack sufficient proof to directly connect these actors to state structures.

A **state-sponsored actor** is a designation commonly used by threat-intelligence firms to indicate a strong possibility of government support. This conclusion typically relies on a synthesis of technical data and open-source research.

State-integrated actors have moved beyond the realm of strong suspicion and have been formally implicated in legal indictments by other countries or named in official statements, linking them to specific government agencies. Such classification may enable a government to use additional mechanisms such as sanctions or call upon alliance commitments.

Finally, a **state actor** openly conducts operations on behalf of a government and publicly acknowledges its affiliation, removing the need for further attribution analysis.

The framework adds an additional layer of resolution to the IISS Cyber Power Matrix platform, which collates examples of state cyber power from 2001 onwards, including

[44] International Institute for Strategic Studies (2024-11-20) *The six degrees of cyber attribution*. https://www.iiss.org/cyber-power-matrix/the-six-degrees-of-cyber-attribution/

state-linked cyber- and influence operations, and the ownership, supply, and physical disruptions of submarine cables.[45]

3.6 Debunk.org

Another relevant actor is the independent technology think tank and NGO Debunk.org. They operate in the Baltic countries, Poland, Georgia, Montenegro, the United States, and North Macedonia, collaborating with partners to analyse and counter disinformation.

They highlight that this area of work poses several challenges. Efforts to debunk false or misleading online content are often fragmented, making unified action difficult. Fact-checking is time-consuming, especially when initial points of investigation are unclear. Additionally, the resources required to debunk disinformation are significantly greater than those needed to create it, and disproving false claims in real time is particularly challenging. These issues highlight the need for comprehensive solutions to effectively counter disinformation.

Informed by this perspective Debunk leverages the expertise of a multidisciplinary team of analysts and works closely with national institutions in partner countries to gain valuable regional insights. They also employ information technology specialists proficient in artificial intelligence tools to streamline and enhance the fact-checking process. Furthermore, a community of volunteer fact-checkers in Lithuania supports the organization's undertaking to expose and counter disinformation.

Debunk state that they employ a comprehensive methodology for detecting and debunking disinformation, combining advanced artificial intelligence technologies with the expertise of analysts and fact-checkers.[46] Before labelling any content as disinformation or misinformation, the organization follows a three-step procedure:

1. Source Identification (Who?)
Analysts begin by assessing the credibility of the source. They examine the most frequent issues addressed by the author, website, or media outlet to determine their reliability and trustworthiness. This step also involves checking the recency of the source's publications to ensure relevance and consistency.

2. Content Assessment (How?)
The next phase involves a critical evaluation of the content itself. Analysts look for suspicious elements in photos, quotes, interviews, or posts. They assess whether headlines are sensational or emotionally charged and verify if the article's text supports

[45] *The IISS Cyber Power Matrix*. https://www.iiss.org/cyber-power-matrix/overview/

[46] Debunk.org. (n.d.). *Methodology.* Retrieved from https://www.debunk.org/methodology (Accessed: November 27, 2024)

the headline. Additionally, they analyse the author's intended message and the rhetorical techniques used to engage the reader.

3. Circumstance Assessment (When?)
Finally, analysts consider the context surrounding the information's appearance. They investigate the circumstances under which the content was published and check for patterns or similarities with other narratives or sources. This helps identify coordinated disinformation efforts and understand the broader implications.

In addition to this procedure, Debunk.org integrates the framework of the Pillars of Russia's Disinformation and Propaganda Ecosystem, developed by the Global Engagement Center of the U.S. Department of State.[47] This framework outlines five key pillars through which Russian disinformation operates: Official Government Communications, State-Funded Global Messaging, Cultivation of Proxy Sources, Weaponization of Social Media and Cyber-Enabled Disinformation. These pillars represent a spectrum of activities ranging from overt government statements to covert cyber operations, with the difficulty of attributing Russian involvement increasing from the first to the last pillar.

Furthermore, Debunk.org applies the Breakout Scale concept devised by Ben Nimmo of the Atlantic Council's Digital Forensic Research Lab.[48] This scale categorizes influence operations into six categories based on their reach and impact: 1: one platform, no breakout, 2: one platform breakout or multiple platforms, no breakout, 3: multiple platforms, multiple breakouts, 4: cross-medium breakout, 5: celebrity amplification, and 6: policy response or call for violence.

By utilizing the Breakout Scale, researchers can classify and monitor the evolution of influence operations over time. Each category captures the status of an influence operation at a given moment, meaning operations can progress upward or regress downward along the scale. The goal is to support the prioritization of resources and improve the coordination of responses to different disinformation tactics.

Debunk underscores that attributing responsibility in regards to influence operations is a complex process that doesn't rely on simple formulas.[49] Instead, it often involves detecting mistakes made by malign actors and connecting various clues, or breadcrumbs, to identify who is behind different types of information influence efforts. The data required for attribution varies with each investigation. For example, uncovering who operates a channel, an account, or a bot farm requires different

[47] U.S. Department of State, Global Engagement Center. (2020). *GEC Special Report: Pillars of Russia's Disinformation and Propaganda Ecosystem*.

[48] Nimmo, B. (2020). *The Breakout Scale: Measuring the impact of influence operations*. Washington D.C.: The Brookings Institution.

49 Interview with Viktoras Daukšas, Director of Debunk.org, November 27, 2024.

approaches and data sets than analysing accounts on fake financial platforms. The type of data available can also differ depending on the platform and domain provider.

The third step of contextual analysis within the circumstance assessment in Debunks methodology, following source identification and content assessment, is particularly important. When investigating influence operations by known threat actors, analysts examine their new operations and tactics. These actors sometimes spread malinformation, facts presented within false interpretations and contexts, to manipulate audiences while appearing credible.

By collaborating with other projects and codifying Foreign Information Manipulation and Interference (FIMI) cases, Debunk continuously adds to their comprehensive knowledge base of threat actors and Indicators of Manipulation (IOMs). Each investigative report contributes specific IOMs that can be linked to particular threat actors. As more reports are compiled, a greater number of IOMs are accumulated, which provides valuable context for analysing new influence activities. This allows for more professional attribution of operations from multiple perspectives. Analysts are assigned to examine each observable element to determine how they relate. By building FIMI case datasets using frameworks like STIX and integrating them with a Cyber Threat Intelligence Platform, these datasets can be correlated with observables to identify IOMs effectively. They also collaborate with other networks and think tanks to consolidate findings into joint investigations or to obtain independent verification of their analyses. This cooperative approach enhances the credibility of their findings and contributes to a more comprehensive understanding of influence operations.

Unlike political institutions that might use attribution as a form of strategic communication, these researchers focus on providing detailed information to others so they can independently confirm or refute the attribution findings. Organizations such as Debunk can submit their incident and investigation reports to alliances or government bodies. If these entities find a report compelling, they may allocate resources to verify the findings and proceed with formal attribution. Independent actors also serve as trusted flaggers on certain social media platforms and can submit high-priority reports of actions that violates the platform's policies or legal regulations.

If platforms fail to take appropriate action, these actors can escalate the issue to the Digital Services Coordinators. These coordinators are responsible for overseeing the application and enforcement of the EU Digital Services Act in their respective member states, ensuring that platforms adequately protect EU citizens.

3.7 DFRLab

The Digital Forensic Research Lab (DFRLab) established in 2016 as part of the Atlantic Council, focuses on studying and producing guidance in terms of methods and strategies on how to address the information environment and the wider information ecosystem including the technologies that influence it. With a team of over 30 experts based on five continents, the DFRLab works on issues like disinformation, online harms, and foreign interference through research, real-time reporting, and training programs. Its goal is to counter disinformation, support democratic systems, and strengthen digital resilience around the world.

They are organized around three main activities. First, conducting open source research, providing analysis on topics such as false information, platform policies, foreign influence, and how the information ecosystem functions. This research informs responses to information challenges and serves as a resource for governments, civil society, and the private sector. Second, it sets standards for research methodologies and trains individuals and groups worldwide in how to practically apply these techniques. These training efforts enable others to monitor and address digital threats locally and to integrate knowledge of the digital environment into their work. Third, the DFRLab uses its findings to create policy recommendations and build partnerships.[50]

In a recent publication, the DFRLab outlines the methodology for selecting cases and implementing a scoring system for its Foreign Interference Attribution Tracker (FIAT), launched on October 23, 2024.[51] This open-source database documents allegations of foreign malign influence and interference in the November 2024 U.S. general election, assessing their credibility, transparency, and broader impact on political discourse.

FIAT's release comes amidst heightened concerns over foreign interference, despite U.S. government efforts, such as warnings and enforcement actions by agencies like the Office of the Director of National Intelligence and the Department of Justice. By its launch, FIAT had documented over 40 cases since the November 2022 midterms. These allegations vary in sources, methods, and objectivity, creating challenges for policymakers, journalists, and others seeking to understand and address these activities. FIAT aims to provide an independent record, establish public attribution standards, and bolster resilience against future interference, especially in the digital sphere.

FIAT 2024 builds on its 2020 predecessor with continuous updates and improved methodologies. Initial findings reveal that nearly half of the documented cases involve interference from China, Iran, and Russia, employing digital tactics like creating fake social media accounts to amplify divisive narratives. Regional patterns have also

[50] DFRLab. (n.d.). *About DFRLab.* Retrieved from https://dfrlab.org/about/ (Accessed: December 3, 2024)

51 Sadek, D, Furbish, M. & Rizzuto, M. (2024) *DFRLab launches the 2024 Foreign Interference Attribution Tracker*, Digital Forensic Research Lab (DFRLab), https://dfrlab.org/2024/10/23/dfrlab-launches-fiat-2024/

emerged: China has propagated conspiracy theories and divisive content, for example through its "Spamouflage" campaign and Iran has conducted cyber operations targeting presidential campaigns as well as disseminated disinformation tied to the Middle East conflict. Meanwhile, Russian actors have focused on undermining confidence in U.S. elections and democracy while spreading narratives against support for Ukraine. New actors, such as Indian and Israeli entities, have also targeted U.S. audiences on domestic and international political issues.

FIAT focuses on cases involving allegations of digital interference directly related to the 2024 election, excluding routine state media output and concentrating instead on targeted campaigns. To evaluate cases, it employs two analytical frameworks: the Breakout Scale and the Attribution Score. The Breakout Scale categorizes cases into six tiers based on their reach and impact, from isolated incidents to those triggering policy responses or violence.[5253] The Attribution Score assesses claims using 18 binary criteria across four subcategories: credibility, objectivity, evidence, and transparency. These criteria evaluate the impartiality of sources, the avoidance of bias, clarity of evidence, and openness of attribution processes.

Another recently released report from DFRLab details methodological insights gained from a collaborative investigation with NetLab UFRJ.[5455] This joint analysis, conducted between June 1 and August 15, 2024 focused on identifying potential instances of deceptive AI use during the pre-campaign period for Brazil's municipal elections, where some 150 million people vote to elect mayors, councillors, and other local officials across 5,568 municipalities.[56]

The investigation focused on publicly accessible content shared on social networks, messaging platforms, and websites. Recognizing that producers of deepfakes intended for disinformation or harm are unlikely to label their content as such, researchers relied on user discussions or reports referencing political deepfakes. These reports often form the basis for complaints submitted to Brazil's Superior Electoral Court (TSE). However, this approach reveals significant limitations: the most sophisticated deepfakes may not be recognizable as fake to ordinary users, reducing the likelihood

52 The Breakout Scale is described in more detail in the previous section presenting Debunk.org

53 Nimmo, B. (2020) *The Breakout Scale: Measuring the impact of influence operations*. Washington D.C.: The Brookings Institution

54 NetLab UFRJ. (n.d.). *About NetLab UFRJ.* Federal University of Rio de Janeiro, School of Communication. Retrieved from https://netlab.eco.ufrj.br/en/sobre (Accessed: December 3, 2024)

55 NetLab UFRJ, located within the School of Communication at the Federal University of Rio de Janeiro, is a research lab specializing in internet and social network studies. They investigate digital disinformation and its social repercussions in Brazil.

56 Farrugia, B. (2024). *The Challenges of Identifying Deepfakes Ahead of the 2024 Brazil Election.* Digital Forensic Research Lab (DFRLab). Retrieved from https://dfrlab.org/2024/10/02/brazil-election-ai-research

of reporting, while discussions about deceptive AI may fail to include relevant keywords, further complicating detection efforts.

The study monitored election-related conversations on platforms including X, Facebook, Instagram, and YouTube, using the tool Junkipedia. Researchers identified 815 official profiles of mayoral pre-candidates from state capitals, analysing 58,271 threads posted by these candidates. Using Portuguese equivalents for keywords such as "deepfake" and "artificial intelligence," the search returned few relevant results. One notable example was a video posted on July 1 by federal deputy and São Paulo mayoral pre-candidate Kim Kataguiri. The video, acknowledged by the candidate as AI-generated, humorously depicted President Luiz Inácio Lula da Silva fleeing a fictional scenario. While the post complied with Brazil's regulatory requirement to disclose AI-generated content, the humour-oriented nature of the video underscores the challenge of categorizing such content within regulatory frameworks.

In addition to social media, the research examined messaging platforms WhatsApp and Telegram. On WhatsApp, 1,588 public groups comprising nearly 48,000 subscribers were analysed, while Telegram yielded 854 groups with over 76,000 subscribers. Searches for keywords related to deepfakes or AI-generated content did not reveal any relevant electoral deepfakes, and expanded queries focusing on phrases like "AI-generated" or "manipulated by AI" similarly returned no results.

To investigate political advertising, researchers analysed Meta’s Ad Library, which hosts promotional content from Facebook, Instagram, and Messenger. Despite Meta’s policy requiring labels for AI-generated content, the platform lacks functionality to search for such labels, creating challenges for researchers. Previous cases from European elections, where Meta’s enforcement was inconsistent, highlight the difficulty of reliably identifying AI content through existing tools.

To complement these methodologies, the team used Google Alerts to monitor blogs and news websites for mentions of relevant keywords. This approach yielded reports of 16 deepfake cases from various Brazilian states, with most incidents originating from WhatsApp groups. In several cases, victims pursued legal action to get the deepfakes removed and seek out the perpetrators.

One notable incident occurred in the city of Igarapé do Meio, where a manipulated video falsely implicated a local mayor in financial fraud. Another in São Paulo involved a fake voiceover impersonating President Lula, disseminated to mislead voters about an electoral event. These cases demonstrate how deepfakes can be weaponized for reputational harm, often requiring victims to take extensive measures to counteract their impact.

The research highlights significant methodological and technical challenges in monitoring AI-generated content. Keyword searches often yielded irrelevant results or false positives, while platform limitations hindered the ability to systematically identify

and track AI-generated material. The Google Alerts methodology, while somewhat effective, was limited to publicly documented cases, underscoring the reliance on external reporting for identifying deceptive content.

Moreover, the absence of robust AI detection tools complicates efforts to identify manipulated content at scale. Current tools, while promising, are prone to inaccuracies, risking false positives or negatives that could mislead researchers and the public. The manual nature of current detection efforts further underscores the need for improved platform functionalities and tools.

Despite these challenges, the report highlights the potential for leveraging public participation in identifying deceptive AI. Fact-checking organizations in Brazil, such as Agência Lupa, already invite the public to submit questionable content via platforms like WhatsApp for verification. Expanding such initiatives to focus on deepfakes and AI-generated content could enhance detection efforts during critical electoral periods.

3.8 Bellingcat

Bellingcat is an independent investigative collective consisting of researchers, investigators, and citizen journalists united by a focus on open source research. Established in 2014, they currently have over 30 staff and contributors across more than 20 countries. The organization has been an early adopter and developer of digital open source collection and analysis methodology to investigate a diverse array of subjects of public interest. These investigations have encompassed events such as the shooting down of flight MH17 over eastern Ukraine, instances of police violence in Colombia, and the illegal wildlife trade in the United Arab Emirates. Bellingcat describe their work as being at the convergence of advanced technology, forensic research, journalism, transparency, and accountability.

The work produced by Bellingcat is frequently referenced by international media outlets and has been cited by several courts and investigative missions. The organization designs and disseminates verifiable methods of ethical digital investigation, publishing walkthroughs of open source research techniques and providing tailored training sessions for journalists, human rights activists, and members of the public.[57]

In 2022, Bellingcat and the Global Legal Action Network released a report introducing a methodology for online open source investigations, based on their work analysing incidents occurring in Ukraine after the Russian invasion. The document serves as a guide for Bellingcat's investigators conducting formalized investigations, specifically within the context of their Justice and Accountability Unit’s work in Ukraine. It provides practical, step-by-step procedures to standardize investigative practices while

[57] Bellingcat. (n.d.). *Who We Are.* Retrieved from https://www.bellingcat.com/about/who-we-are/ (Accessed: December 1, 2024)

complementing broader principles outlined in frameworks like the Berkeley Protocol on Digital Open Source Investigations. The Berkeley Protocol offers high-level guidance, and this methodology translates those principles into actionable steps tailored to specific investigative goals.

Designed for organizations with adequate resources, the methodology emphasizes secure digital infrastructure, dedicated devices, and thorough investigative practices, which may not suit operations focused on rapid data analysis. Rather than teaching online investigation techniques like locating or verifying content, it assumes investigators are already skilled in these areas and focuses instead on the surrounding aspects, such as ensuring legal admissibility of evidence. It is intended to serve as a comprehensive operational framework for producing high-quality, legally sound investigations.[58] Worth bearing in mind is that any specific tools mentioned in the methodology report were recommended for a specific investigation in 2022, and may have changed since.

The methodology is comprised of eight phases:

1: Systems and resources
Investigators are required to use a dedicated work device exclusively for their investigations, ensuring that all data is secure through encryption and strong passwords. To protect their online activities and maintain privacy, a virtual private network (VPN) is supplied. Communication within the team is facilitated through Slack, allowing for efficient collaboration and information sharing.

Investigators are given access to cloud storage to organize and store their research materials systematically. Within the Google Chrome browser, tools like Hunchly that helps save and store copies of websites, or online content exactly as they appear at a certain moment, are used to track and preserve online activities. Meanwhile, Google Sheets and Documents help in managing and documenting data, research notes, and incident assessments. Additionally, Uwazi, a web-based database application developed for human rights defenders to gather and organize collected information, is utilized in a separate browser window to analyse and manage information related to incidents, media content, and involved parties.

Overall, these guidelines ensure that investigations are conducted securely, systematically, and collaboratively, leveraging specialized tools to maintain thorough and organized records throughout the investigative process.

2: Briefings
Investigators are briefed on international humanitarian and criminal law to understand the evidentiary standards and legal principles relevant to their investigations, focusing on documenting incidents neutrally and thoroughly. They are not tasked with making

[58] Bellingcat & Global Legal Action Network. (2022). Methodology for Online Open Source Investigations into Incidents Taking Place in Ukraine Since 24 February 2022.

legal judgments but instead produce detailed analyses and summaries to support potential legal assessments.

Practically, investigators follow a style guide and an incident template to ensure consistent and professional reporting. They should also collaborate to build shared resources, like databases on weapons available to different actors, their identifying characteristics and effects, in order to enhance the investigators' ability to effectively analyse evidence.

3: Categories of information

Open source investigators should categorize and cite the information they collect. Sources are divided into two main types: examinable content and descriptive content. Examinable content includes granular items such as videos, photos, satellite imagery, social media posts, and tools like aviation or maritime trackers. These can be directly analysed and examined to verify their authenticity, cross-reference with other sources, and draw conclusions. This type of content forms the core of an investigator's work as it allows for detailed examination and verification.

Descriptive content, on the other hand, is content published online, usually in written form, that describes events but falls outside the scope of analysis using online open source investigation techniques. This includes written accounts such as news articles or NGO reports that describe events. While descriptive content can provide reliable context and valuable information, it is not the primary focus of open source investigations. The distinction helps investigators focus on material that can be scrutinized and pieced together while still acknowledging the importance of descriptive content for supporting their findings.

4: Preparation

Investigators are required to secure their devices with encryption, strong passwords, and separate user accounts for their investigative work. Essential tools and a VPN must be installed to protect the investigator's identity and ensure secure internet access. The VPN's settings are specific, requiring exit nodes to be set in such a way that it can bypass content restrictions, with any deviations approved by the lead investigator.

To maintain anonymity and access restricted information, investigators create virtual identities on social media platforms like Facebook and Twitter, while ensuring personal accounts are logged out on platforms like YouTube and Instagram. These virtual accounts also help reduce algorithmic biases that could affect search results. Investigators must record all research accounts and related details in a designated tracking document.

The process includes training search engine algorithms by performing searches in relevant languages (English, Ukrainian, and Russian) to tailor results to the investigation's needs. Direct interaction with sources is prohibited without prior authorization, ensuring ethical and consistent communication practices.

Team collaboration is critical, facilitated through Slack for general discussions and Signal for handling sensitive topics. Investigators must use designated communication channels, avoiding personal or unrelated threads, to ensure professionalism and data security throughout the investigation.

5: Investigation

Investigators must use a VPN, log out of personal accounts and log into dedicated research accounts for accessing necessary platforms securely. To organize the investigation, a specific folder structure is created in a shared cloud system, ensuring all related documents, links, files, and notes are systematically stored using predefined templates.

Since all web activity related to the case is tracked, investigators need to carefully manage their browsing to ensure personal activities are kept separate from investigative work, either by using a different browser or documenting exceptions transparently. This prevents contamination of investigation records while ensuring privacy.

There are also protocols for accessing restricted content, like graphic videos, by using verified accounts under controlled circumstances. These steps are to ensure that the investigation is fully auditable, allowing for transparency and accuracy in the findings.

6: Discovery / Content Gathering

In this phase the key focus is on maintaining objectivity, organizing evidence, and preserving content for analysis and future use. Before starting, investigators are encouraged to review legal guidelines and familiarize themselves with international humanitarian law to ensure they recognize relevant information. They should plan their investigation using for example a Table of Factual Inquiries, which outlines the critical aspects of an incident or case that require verification or analysis, as well as keep an open mind to explore all possibilities and avoid cognitive biases.

During the investigation, every search term and decision must be logged in a Research Notes document to ensure the process is replicable and traceable. In addition to the tool archiving web activity, the content itself is preserved by Mnemonic, a system that stores videos, images, and other media for long-term use. All content must be organized into a structured folder system and cited accurately, including identifying the original online source whenever possible.

Examinable content is analysed and logged in the database used for evidence organization and case-building. Investigators must also download and save relevant media files with specific naming conventions in their private cloud folders, ensuring consistency and easy reference. If any content may benefit another team member, it should be shared via the designated communication channels and transferred to a shared folder.

For satellite imagery, while preservation isn’t required due to low risk of removal, it can still be downloaded for reference and grouped in the content database for analysis.

Throughout the process, investigators collaborate with the legal team to ensure all evidence is logged correctly, analysed thoroughly, and integrated into case files for future review.

7: Verification & Analysis

During the investigation, the goal is to use methods like geolocation, chronolocation, corroboration, and cross-referencing to verify content. These findings are documented in the Incident Assessment Report, a living document where investigators neutrally describe their verification work and its significance, using accessible language for an informed audience. While detailed geolocation reports might be required later, this stage focuses on summarizing findings effectively.

When reporting deaths or casualties, investigators are advised to use cautious, neutral language, avoiding assumptions about civilian or combatant status unless supported by direct evidence. Numbers of casualties are only confirmed when examinable items, such as official documents or media footage, can verify them. Reports from external organizations are included as descriptive content but not analysed unless they can be cross-verified. Investigators are guided to use clear, confident language when drawing conclusions but must avoid overstating certainty unless there is strong evidence. For example, they might use terms like “appears to be” or “suggests” but avoid unnecessary hedging if the evidence allows a firm conclusion.

At the end of a session, investigators securely log their findings by saving links, updating archives, and uploading relevant content to shared folders. They ensure that all temporary files are deleted, browsers are closed, and the device is powered off to minimize security risks. This process ensures the investigation is comprehensive, well-documented, and adheres to security and professional standards.

8: Roles and responsibilities

This phase outlines the structure of the team and related experts that the investigator needs to be aware of. This includes the Lead Investigator who oversees the process, First Instance Investigators who conduct initial research, and Reviewing Investigators who verify findings. Coordination with the legal team is emphasized, including Lead Lawyers and Ad-hoc Legal Consultants, who help design procedures and ensure compliance with legal standards.

The Bellingcat open source investigation methodology offers a useful example of a practical guide to navigating the complexities of using digital evidence in accountability processes. By bridging investigative practices with evidentiary principles, based on real-world experience, it reinforces the need for cross-skillset collaboration and emphasizes the importance of rigour, transparency and security when collecting and analysing data during conflict and war.

3.9 The Security, Crime, and Intelligence Innovation Institute

The Security, Crime, and Intelligence Innovation Institute at Cardiff University has through their recent work provided a good example of how to synthesize a case study based on available open source data from other actors. They conducted an extensive investigation into the Ghostwriter cyber-enabled influence campaign, widely attributed to actors linked with the Russian state[59]. Active from 2016 to 2021, the campaign combined hacking techniques with disinformation strategies, targeting multiple countries and continually adapting its methods to exploit various digital platforms. The study analysed 34 incidents documented by cybersecurity firm FireEye/Mandiant[60], supplemented with data from government communications, media reports, fact-checking organizations, and think tanks. To enrich their findings, researchers conducted nine in-depth interviews with government officials, media representatives, and civil society members involved in exposing or countering Ghostwriter's activities.

The investigation developed a comprehensive timeline of incidents, illustrating the campaign's evolution and its capacity to escalate operations. Early tactics were straightforward, such as website compromises and impersonations, reflecting core motivations and confidence in these methods. Over time, Ghostwriter adopted more sophisticated techniques, including longer distribution chains, simultaneous multi-country targeting, and the use of new platforms like Telegram. Its activities extended beyond the analysed period into countries like Belarus, Germany, Lithuania, Poland, and Ukraine, underscoring its persistent threat.

Key research questions guided the study, focusing on how the campaign evolved, why it circumvented countermeasures, and what implications it holds for understanding modern disinformation efforts. The methodology involved coding incidents based on targeted countries or organizations, languages used, and specific influence tactics such as fake emails, fabricated websites, manipulated media, and falsified statements. Tools like CrowdTangle[61] were employed to assess the presence of Ghostwriter's content on platforms like Facebook. However, limitations arose due to the inability to trace some original messages and responses, highlighting challenges in relying solely on open source data.

Despite these limitations, the collected data provided valuable insights into Ghostwriter's distinctive operational patterns. Attribution of the campaign has been fragmented and complex, with different entities—including private cybersecurity firms, governments, and international organizations—linking aspects of Ghostwriter to

[59] Cardiff University Security, Crime, and Intelligence Innovation Institute. (2023). The Ghostwriter Campaign as a Multi-Vector Information Operation: Attempts to Control Its Influence & the Limitations of Current Counter-Measures.

[60] Mandiant. (2021). Ghostwriter update: cyber espionage group UNC1151 likely conducts ghostwriter influence activity.

[61] CrowdTangle was a tool from Facebook, now Meta, that was used to explore public content on social media. It was discontinued on August 14, 2024

Russian military intelligence or the Russian state, and in some cases, to Belarusian actors. Public knowledge about the exact operators behind Ghostwriter's content creation remains limited. Russian-language media have strategically amplified rebuttals to Ghostwriter's incidents while mocking Western responses, adding complexity to attribution efforts.

The IIO Attribution framework was also used to aid in understanding differences between the actors that have attributed Ghostwriter. Three main discrepancies and weaknesses of the current system for countering information operations were detected:

Classified Government Data: Much evidence connecting Ghostwriter to state actors is classified, restricting public understanding and accountability. For instance, Germany and the EU have made statements but often lack detailed evidence or fail to specify affected member states. Even though most of the public callouts appear framed to act as a deterrence they had not been followed by any further measures despite the campaign continuing to pose a threat.

Dependence on Proprietary Sources: Private sector actors such as cyber defence and threat intelligence company Mandiant have raised awareness but often rely on undisclosed methodologies or data. This requires users and the general public to rely on their attribution and potential countermeasures taken without much detail or independent validation. There are exceptions, of course, as when Meta documented some of their countermeasures against Ghostwriter influence activities in an April 2022 Adversarial Threat Report.[62]

Limitations of Open Source Data: Ghostwriter has obscured its origins by using the Russian language and platforms, while Russian State media avoided directly promoting or disseminating campaign related messaging. This has reduced the effectiveness of open source analyses and left gaps in the understanding of the campaign's full scope.

The orchestrators of the campaign capitalized on these vulnerabilities, adapting its methods to evade detection and mitigate the impact of countermeasures. The study underscores the campaign's resilience, especially its shift from targeting individual countries to orchestrating simultaneous, multi-national operations. By 2021, its tactics had become more sophisticated, involving extended distribution chains and moving to new dissemination platforms. These included leveraging hacked social media accounts on Twitter and Facebook as well as the utilization of growing platforms like Telegram to broaden its reach.

The Cardiff University analysis offers insights into the complexities of investigating and conducting comprehensive attribution of persistent, multimodal cyber-enabled influence campaigns. It emphasizes the need for greater transparency in attribution processes, enhanced collaboration between governments and private entities, and more robust open source strategies to combat disinformation. Ghostwriter exemplifies the

[62] Nimmo, B., Agranovich D. & Gleicher, N. (April 2022). *Adversarial Threat Report*. Meta.

challenges posed by modern information influence operations, where well-resourced malign actors can leverage adaptability, obscured origins, and multi-platform amplification.

3.10 Protection Group International

A private sector organization providing a wide portfolio of services related to managing digital risks is Protection Group International (PGI), established in 2013. Their work includes capacity building, crafting solutions to enable governments and organizations to develop their own cyber and open source intelligence capabilities. In digital security, PGI provides services to implement or support the implementation of digital security controls and incident response mechanisms. The company also conducts digital investigations, offering services to identify, mitigate, and effectively respond to malign content on social media.[63]

Due to the sensitive nature of their work, PGI does not disclose specific capabilities and methodologies in detail. However, the structure of the companies range of solutions can give an insight into approaches for- as well as the interdisciplinary nature of addressing complex threats. Their service offering is organized into three main specializations which they refer to as **Detect, Protect, and Build**.[64]

The **Detect** team, likely most related to attribution-related activities, focuses on identifying and understanding online threats to provide actionable insights, tracking threat activities from actors that employ and exploit digital spaces for their purposes. By understanding the evolution of these threats, it anticipates future tactics and activities, aiming to reduce the impact of adversarial groups on companies and citizens. The process involves initiating tracking through adaptable methodologies, observing and analysing threat activities, and identifying ways to mitigate the influence of such actors, including assisting in the removal of harmful networks from social media platforms.

PGI also states that they identify, analyse, and disrupt new forms of digital threats as they emerge. Recognizing that adversarial actors continually adapt to overcome security measures, it employs a combination of technology and human expertise to detect behaviours such as doxing, coordinated harassment, mass reporting, and other malign activities. By mapping and assessing evolving threats, identifying and disrupting them at their source, and advising on policy enforcement gaps, thereby aiding platforms and regulators in maintaining a proactive defence.

---

[63] Global Forum on Cyber Expertise. (n.d.). *Protection Group International (PGI).* Retrieved from https://thegfce.org/member-and-partner/protection-group-international-pgi/ (Accessed: December 1, 2024)

[64] Protection Group International. (n.d.). *About Us.* Retrieved from https://www.pgitl.com/ (Accessed: December 1, 2024)

In addressing national information integrity, PGI employs a holistic approach to building a resilient information environment. By analysing the entire information supply chain—from content creation to distribution and consumption—they work to raise clients understanding how information is used to undermine critical national processes and services. This work includes mapping information sub-communities, tracking early signs of malign activities, and building infrastructure for information sharing to aid in the early identification of malicious campaigns.

In regions affected by conflict, PGI identifies, maps, and attributes digital activities and campaigns run by hostile actors such as foreign states and violent extremists highlighting how these actors exploit digital capabilities to destabilize power structures and disrupt peace processes. PGI also supports clients by developing investigative skills among employees, assisting in the creation of social media codes of conduct, and helping to establish public education and awareness campaigns about hostile influence operations, disinformation, and hate speech.

In the **Protect** team, PGI focuses on enhancing clients' security measures through several services. It conducts security testing to identify vulnerabilities relevant to the specific threats that clients face, helping them understand the necessary steps to address and remediate these issues. The organization provides incident response services that include both proactive and reactive components. Proactively, PGI assists in preventing incidents by strengthening security measures and preparing organizations to handle potential threats; reactively, it offers emergency response services to manage and mitigate the impact of security incidents that have occurred. They also aid in developing information security management systems and assists clients in achieving compliance with international and national standards, ensuring that organizations maintain robust information security practices aligned with recognized frameworks.

The **Build** team is dedicated to developing long-term capabilities and infrastructures to support national and organizational digital resilience. According to the company, they draw on experience from over 30 national capability development projects in more than 80 countries, offering expertise across various national priority areas. They collaborate with nation-states, multilateral organizations, and NGOs to provide subject matter expertise, assisting in deploying funding and resources effectively, especially in situations where leaders face additional responsibilities and operate amid a global shortage of skilled personnel. They also design future-focused workforce frameworks to support nation states define and prioritize their requirements to enable efficient capability building and an optimization of limited resources. In addition, they assist in developing curricula, career pathways, and accreditation processes for cybersecurity professionals. Other related services include identifying knowledge gaps within organizations, conducting training needs analyses, and assisting in the design and establishment of Security Operations Centres (SOCs), Cyber Security Incident Response Teams (CSIRTs) and social media monitoring desks.

4 How the framework has been used

Since its introduction in 2022, this initial iteration of the framework has been adopted by organizations such as Microsoft,[65] the EEAS StratCom Division, Recorded Future,[66] CSCIS, and at least one intelligence agency. Notably, Microsoft's Digital Threat Analysis Center (DTAC) used the matrix to outline the key types of evidence underpinning their attribution of an Iranian influence operation.[67]

4.1 Microsoft

Microsoft has adopted and modified the Attribution Framework to incorporate it into their Digital Threat Analysis Center (DTAC) operations to work on complex influence operations. The DTAC acknowledges that publicly available standards for attributing influence operations remain largely undeveloped. To address this gap, they have adopted a modified version of the Pamment and Smith (2022) framework, which considers three categories of evidence, technical, behavioural, and contextual either in the open source alone or enriched with proprietary[68]. The technical evidence includes investigations into domain ownership, IP address ranges, and financial transactions. Open source analysis draws on publicly accessible resources such as domain registration details and corporate registries. This is complemented by proprietary sources, including telemetry data from platforms, which provide additional insights such as account creation dates, email addresses, and platform usage habits. These combined data sources add significant depth to the analysis. However, the use of advanced masking techniques, such as proxy servers and VPNs, often complicates the process and demands further analysis. Nevertheless, the attributions made by DTAC are almost always based solely on open source evidence

Behavioural evidence focuses on the observable actions and patterns of accounts or entities suspected of conducting influence operations. For open source, publicly hosted data, this involves analysing posting habits, cross-platform sharing, amplification techniques, and engagement within social networks. Microsoft refines these analyses with proprietary evidence, allowing for deeper insights into activities within private groups and identifying countermeasures employed by actors to obscure their

[65] Microsoft Threat Analysis Center. (2023). *An attribution model for influence operations,* p. 2. Microsoft.

[66] Recorded Future, Insikt Group. (2024). *"Operation Undercut" Shows Multifaceted Nature of SDA's Influence Operations*, p. 7. Recorded Future.

[67] Microsoft. (2023). *DTAC: Charlie Hebdo Hack, Iran, and Neptunium.* Microsoft.

[68] Microsoft (2023). *MTAC Attribution Model for Influence Operations*. Microsoft.

operations. This behavioural analysis helps establish patterns of operational consistency or detect anomalies that may indicate coordinated efforts behind a operation.

Contextual evidence builds on narrative analysis and the geopolitical environment in which influence operations are taking place. Open source contextual analysis maps linguistic markers, political motives, and the narratives being promoted. Proprietary evidence enhances this process by identifying patterns from previous campaigns and linking them to current operations. This evidence is particularly useful for understanding how disinformation targets specific audiences and aligns with the strategic objectives of state or non-state actors.

To communicate the certainty of their findings, DTAC employs estimative language to articulate confidence levels ranging from low to high. These assessments depend on the quality and corroboration of the evidence and align with standards set by the U.S. intelligence community. By using probabilistic language such as "likely" or "assess," DTAC ensures that their conclusions remain nuanced and that they avoid unwarranted impressions of certainty.

	Technical evidence	Behavioral evidence	Contextual evidence
Open source	Domain ownership, IP ranges, documented financial relationships, etc.	Account or page activity, posting patterns, cross-posting, sharing patterns, social network analysis	Political context, narrative analysis, analysis of media, linguistic markers, possible beneficiaries
Proprietary source	Data sourced through proprietary telemetry or platform backend	As with open source, enriched by proprietary platform data	As with open source, enriched by proprietary data from previous attributions and disclosures

Figure 3: DTAC's Influence Attribution Matrix [69]

In practice, DTAC used its Attribution Model for Influence Operations to attribute the January 2023 cyberattack on Charlie Hebdo to the Iranian actor known as NEPTUNIUM, also identified as Emennet Pasargad. The analysis integrated technical, behavioural, and contextual evidence to substantiate the attribution with a moderate confidence level. This operation combined technical elements, such as data leaks and website defacements, with behavioural indicators, including the use of sockpuppet

[69] Microsoft Threat Analysis Center. (2023). *An attribution model for influence operations,* p. 2. Microsoft.

accounts for amplification, and contextual links to Iranian geopolitical motivations following criticism of the Iranian Supreme Leader by the magazine[70].

The technical evidence in this case focused on the infrastructure and tools used during the cyberattack. The group calling itself "Holy Souls" claimed responsibility for hacking into Charlie Hebdo's database, allegedly accessing personal data of over 200,000 subscribers. This claim was supported by a sample data leak, which French outlet Le Monde verified as authentic. Microsoft observed that the group employed advanced methods typical of Iranian state-linked cyber campaigns, including data exfiltration and website defacement. However, the analysis revealed no direct technical links between the IP address associated with the group's domain and other known Iranian state actors. The domain used by the actors was resolved to an IP address in Malaysia and was not co-hosted with any other suspected Iranian actor sites. Despite this limitation, the tools and methods used in the attack closely aligned with those seen in prior Iranian operations, supporting an indirect attribution to Iranian cyber capabilities.

The behavioural evidence focused on the tactics and online patterns displayed during the operation. After Holy Souls shared the sample data on YouTube and various hacker forums, the leak was strategically amplified across multiple social media platforms. This coordinated effort employed a distinct set of influence tactics, techniques, and procedures (TTPs) that DTAC had previously observed in Iranian hack-and-leak operations. The campaign targeting utilized dozens of French-language sock puppet accounts to amplify its reach and spread antagonistic messages. Starting on January 4, these accounts, many of which had low follower counts and were recently created, began posting criticisms of the Khamenei cartoons on Twitter. Notably, before any significant media coverage of the alleged cyberattack had emerged, these accounts shared identical screenshots of a defaced website displaying the French-language message: "Charlie Hebdo a été piraté" ("Charlie Hebdo was hacked"). A few hours later, at least two additional accounts joined the operation, impersonating French authority figures, one posing as a tech executive and the other as a Charlie Hebdo editor. Both accounts, created in December 2022 and with similarly low follower counts, began sharing screenshots of the leaked Charlie Hebdo subscriber data allegedly obtained by Holy Souls. This added another layer of credibility to the campaign's narrative and mirrored tactics from past Iranian campaigns, including impersonation of credible sources such as journalists and government officials. The accounts also strategically engaged with journalists and news outlets to enhance the operation's visibility. The coordinated nature of this behaviour matched patterns previously observed in Iran-linked campaigns which further supported the attribution. The use of sock puppet accounts is a tactic that has been observed in other Iran-linked

[70] Watts, C. (2023). *Iran Responsible for Charlie Hebdo Attacks*. Microsoft Digital Threat Analysis Center.

operations, such as the 2022 attack claimed by the Atlas Group, a collaborator with Hackers of Savior. This operation was attributed to Iran by the FBI.

The contextual evidence highlighted geopolitical motivations underlying the attack. For instance, the timing coincided with Charlie Hebdo's announcement of a cartoon contest mocking Iranian Supreme Leader Ali Khamenei. This contest drew sharp criticism from the Iranian government, including public condemnations and retaliatory measures, such as summoning the French ambassador and closing the French Institute for Research in Iran. These events provided a strong geopolitical backdrop for the cyberattack. Additionally, the messaging in the operation aligned with Iran's broader strategic goals of silencing critics and intimidating dissenting voices, consistent with historical Iranian state behaviour. The use of French-language sock puppets which was rife with errors indicative of non-native speakers using translation tools demonstrated an attempt to localize the operation and maximize its resonance with French audiences.

	Technical evidence	Behavioral evidence	Contextual evidence
Open source	**No contribution to assessment** • The actor's domain (holysouls[.]cc) has, since January 4, resolved to an IP address in Malaysia (111.90.146[.]101) which is not co-hosted with any other suspected Iranian actor sites	**High confidence** • Use of hacktivist persona + target language, recently created sockpuppet accounts consistent with past campaigns attributed to Iran • Near identical use of reply-post distribution technique aimed at news outlets and journalists consistent with past campaigns attributed to Iran • Networked approach of sockpuppet accounts (i.e., the accounts followed one another) consistent with past campaigns attributed to Iran • Use of "victim-authority" sockpuppet accounts consistent with past campaigns attributed to Iran • Claims of website defacement, data exfiltration, leak of private data online consistent with past campaigns attributed to Iran	**High confidence.** • Iran is actor most likely to want Charlie Hebdo (CH) operations disrupted ahead of Khamenei cartoon issue • Numerous Iranian government official statements (including on same day as release) threatening CH with retribution • Past threats by Iranian government threatening retribution for CH • French language used by sockpuppet accounts is consistent with non-native speakers, suggesting use of translation software • Look and feel of persona accounts consistent with past Iranian government actor personas
Proprietary source	Not used	Not used	Not used
Overall assessment	**High confidence assessment** of Iranian government role in Holy Souls influence operation targeting Charlie Hebdo		

Figure 4: DTAC's Influence Operation Attribution Matrix [71]

[71] Watts, C. (2023). *Iran Responsible for Charlie Hebdo Attacks*. Microsoft Digital Threat Analysis Center.

4.2 Recorded Future

Recorded Future's Insikt Group has employs their "Influence Operation Attribution Framework," which is a tailored adaptation of the original Attribution Framework for analysing information influence operations[72].

The framework was used for the attribution of the Emerald Divide campaign to Iran-aligned actors[73]. The operation consisted of three distinct influence operations carried out sequentially, each with targeted objectives and malign narratives aimed at Israeli society. These operations spanned from the campaign's initiation in 2021 to ongoing activities observed as of February 2024. The first operation sought to amplify social conflict between Israel's ultra-Orthodox religious groups and its LGBTQ+ community. The second operation aimed to incite broader social unrest by exacerbating divisions between individuals on the left and right of Israel's political spectrum. The third and ongoing operation has focused on generating dissatisfaction with the Israeli government's response to Hamas's attacks.

The factors that supported the attribution consist of behavioural evidence, contextual evidence, and overlap with previous findings. The **behavioural evidence** includes activities during the third phase of the Emerald Divide operation, which consistently referenced Iran through both direct and symbolic messaging. A prominent example was a video disseminated by Emerald Divide accounts featuring AI-generated voiceovers and a speculative depiction of Israel in the year 2043. This video prominently displayed the IRGC flag and what appeared to be naval drills conducted by the IRGC Navy. These visual and narrative elements were deliberately crafted to underscore Iran's influence and align the operation with themes typical of Iranian state-backed propaganda. Additionally, the operation circulated posts that included images of Qassem Soleimani, a key figure in Iranian ideological and military narratives, as well as graphics suggesting Iranian arms shipments to the Palestinian Authority.

The **contextual evidence** further supports the attribution of the Emerald Divide campaign to Iranian-aligned actors by highlighting how its narratives align with Iran's strategic priorities. The operation actively promoted divisive narratives designed to intensify ideological tensions within Israeli society by targeting sensitive social and political issues to undermine the Israeli government. These narratives are consistent with Iran's broader goals of projecting regional power and influence while countering Western and Israeli presence through both proxy warfare and "soft war" tactics, which involve the strategic use of information and influence operations. Emerald Divide amplified these narratives by employing advanced digital techniques, including AI-generated imagery, to convey negative sentiment toward key Israeli figures, particularly Prime Minister Benjamin Netanyahu. These AI-generated posts were not

[72] Insikt Group. (2024). *Operation Undercut Shows Multifaceted Nature of SDA's Influence Operations*, p. 7. Recorded Future.

[73] Insikt Group. (2024). *Iran-Aligned Emerald Divide Influence Campaign Evolves to Exploit Israel-Hamas Conflict*. Recorded Future.

only visually striking but also tailored to provoke emotional responses which further aimed to fuel societal divisions and undermining public confidence in Israeli leadership. By leveraging these narratives, the campaign sought to deepen internal discord and align its messaging with Iran's geopolitical objectives of destabilizing its adversaries.

The alignment of Emerald Divide's activities with prior attributions by multiple entities strengthens the case for its connection to Iranian-aligned actors. For example, on February 7, 2024, Microsoft attributed recent activity in the Emerald Divide campaign to Storm-1364, which it assessed as a likely Iranian state-aligned influence threat actor. This attribution aligned with Insikt Group's tracking of Emerald Divide and reinforced earlier findings by the Israel Security Agency (Shin Bet), which had linked specific social media accounts involved in the campaign to Iranian influence operations targeting Israel. The findings also corresponded with reporting from Haaretz on December 20, 2023, which detailed a two-year-long malign influence operation aimed at increasing social and political polarization within Israeli society. This article drew on insights from FakeReporter, an Israel-based disinformation watchdog group, whose research highlighted how these operations were designed to exploit divisions across the Israeli political spectrum. The Haaretz and FakeReporter reports were consistent with Shin Bet's disclosures in June 2023, which outlined Iranian efforts to sow discord through social media. These activities targeted wedge issues within Israel, exacerbating divisions and undermining social cohesion. Shin Bet raised concerns about inauthentic social media accounts attributed to Iran, which were deliberately amplifying polarizing rhetoric to deepen societal divides. This pattern was noted during the country's 2022 parliamentary elections, when Shin Bet cautioned that both Russia and Iran might use divisive narratives and false information to fuel unrest.

In the report, Recorded Future uses DISARM framework to classify specific tactics, techniques, and procedures (TTPs). For example, during the third phase of the campaign, accounts explicitly referenced Iran through visual and symbolic representations. One instance involved a video employing T0097.206: Government Institution Persona, where AI-generated voiceovers narrated a fictitious depiction of Israel in 2043. The video featured imagery of the Islamic Revolutionary Guard Corps (IRGC) flag and what appeared to be IRGC naval drills, reinforcing the narrative that the content was aligned with Iranian state objectives. Such imagery tied the operation's messaging to Iran's geopolitical narrative, portraying its regional power while undermining Israel's stability.

Other examples include the campaign's use of AI-generated videos and images which were disseminated widely to influence perceptions among Israeli audiences, T0087: Develop Video-Based Content, and the use of T0086.003: Deceptively Edit Images (Cheap Fakes) to amplify negative sentiment toward Israeli leadership and crafting an image of incompetence and fostering distrust among citizens.

Another example of how Recorded Future has used the attribution framework is from November 2024 when they exposed an influence operation orchestrated by Russia's Social Design Agency (SDA) that they refer to as "Operation Undercut"[74]. The influence operation has been active since at least December 2023, and it shares thematic and narrative similarities with previous campaigns like Doppelgänger and Operation Overload, yet it operates with separate infrastructure and mechanisms. The campaign employs cutting-edge techniques including AI-enhanced videos and images to impersonate credible Western news outlets and to amplify its messaging via trending hashtags. Notably, Recorded Future identified over 500 social media accounts linked to the operation, with the likelihood of a much larger network, some of which have already been deactivated by platforms. The narratives that are promoted by Operation Undercut aim to discredit Ukraine's political and military leadership, to undermine Western military aid, and to portray the future of this aid as contingent on the results of the 2024 US elections. Beyond Ukraine, Operation Undercut also attempts to amplify socio-political divisions in the United States, leveraging polarizing topics like the Israel-Gaza conflict and the political implications of the 2024 elections. Additionally, the operation seeks to exploit existing tensions in the European Union by criticizing EU leadership and by fostering division between member states and Brussels. Despite its advanced methods, including the use of commercial AI tools like ElevenLabs for realistic voiceovers, the campaign has achieved minimal public engagement and have likely had a limited impact on public opinion. However, Recorded Future highlights the broader risks associated with such operations, including the erosion of trust in media outlets, reputational damage, and financial harm caused by brand impersonation. Recorded Future´s Insikt Group attributes the campaign to the SDA and assess that Operation Undercut is "almost certainly an influence operation conducted by SDA using a network of social media accounts engaging in coordinated inauthentic behaviour (CIB) seeking to achieve influence objectives in support of Russian government interests."[75] Similar to the Emerald Divide campaign, Recorded Future´s Insikt Group applied the framework to attribute this campaign to the SDA.

The attribution made by Recorded Future´s Insikt Group is based on multiple authoritative sources such as OpenAI´s reporting in May 2024 where they identified aspects of the SDA Doppelgänger operation, the U.S Department of Justice (DOJ) affidavit from September 2024 in which they detailed SDA´s involvement in the Doppelgänger campaign alongside Structura and ANO Dialog. Further evidence comes from a September 2024 investigation by media outlets where thousands of leaked SDA documents were uncovered. These files, that were verified by experts and U.S. intelligence, provided critical insights into SDA's coordination with the Russian Presidential Administration, its strategic monitoring of Western socio-political divisions, and its sophisticated content production pipeline involving writers, designers, and editors. By synthesizing the evidence from these sources together with

[74] Insikt Group (2024). *Operation Undercut Shows Multifaceted Nature of SDA's Influence Operations*. Recorded Future.

[75] Ibid., p. 7

evidence from Recorded Future Intelligence Cloud data, the report presents a detailed overview of Operation Undercut´s TTPs and infrastructure.

In their modified version of the attribution framework used for this attribution, **behavioural patterns** and **contextual evidence** served as the key supporting factors for attribution. Behavioural patterns focus on the observable activities and techniques employed by actors, such as account activity, posting frequency, methods of amplification (likes, shares, reports), and interactions within a network. The contextual evidence highlights the narratives disseminated, the cultural and political significance of the content shared, and the underlying motivations of the actors involved.

The **key behavioural patterns** that support Recorded Future´s Insikt Group attribution of Operation Undercut to Russian influence networks include content style, target audiences and languages, and platform manipulation. For example, network analysis revealed high coordination between accounts that shared identical text content, videos, and images. Operation Undercut accounts frequently posted cartoons that shared striking similarities in style and narrative with caricatures found in the leaked Social Design Agency (SDA) documents. This consistency led Recorded Future´s Insikt Group to conclude that "the network's operators very likely have access to in-house content creation capabilities, including graphic designers, video editors, and illustrators."[76] The style and format of these cartoons also bore strong visual similarities to those uploaded by Doppelgänger's inauthentic websites and accounts, further linking Operation Undercut to Russian influence campaigns.

Pattern-of-life (POL) analysis of Operation Undercut´s posting behaviour provided additional behavioural evidence. The analysis showed that the network followed a consistent working schedule, with activity aligning with standard Russian business hours and pausing on Russian holidays. This pattern indicated that the operation was conducted from within Russia. Like Doppelgänger, Operation Undercut accounts posted content in multiple languages, including Russian, German, French, English, Polish, Turkish, and Hebrew. They also boosted engagement by leveraging trending hashtags in the targeted countries.

Despite these similarities, there were notable differences between Operation Undercut and Doppelgänger which suggested that they were distinct operations. For instance, unlike Doppelgänger, which relied heavily on obfuscated links to inauthentic websites and automated accounts for amplification, Operation Undercut produced and distributed a significant amount of unique video and image content directly on social media platforms. These included AI-enhanced videos featuring a combination of human and AI-generated voiceovers and images that impersonated or cited Western news organizations' coverage of Ukraine. According to Recorded Future´s Insikt Group assessment, these uploads were likely managed by human-operated accounts, contrasting with Doppelgänger's reliance on automated systems. Moreover, there was

[76] Ibid., p. 7

no evidence of content sharing or direct collaboration between the two campaigns. Operation Undercut accounts did not post links or materials tied to Doppelgänger websites, nor did Doppelgänger accounts promote content originating from Operation Undercut. The only observed cross-platform engagement involved low-volume promotion of Operation Undercut's 9gag content, which appeared to be managed by human-operated accounts rather than bots.

A central part of the **contextual evidence** supporting Recorded Future´s Insikt Group attribution is the continuity of themes between Operation Undercut and the Doppelgänger campaign. Both operations closely align with objectives and key performance indicators (KPIs) outlined in leaked documents from the Social Design Agency (SDA), as referenced in the DOJ affidavit. This thematic consistency strongly suggests that both campaigns were shaped by similar strategic goals established by the SDA. Between July and September 2024, Operation Undercut employed AI-enhanced videos and images which often impersonated or quoted Western media outlets to propagate narratives aligned with Russian influence objectives. The operation targeted Western military support for Ukraine, seeking to discredit Ukrainian political leadership, accusing Ukraine of financing "terrorist" anti-Wagner forces, and promoting Russia's nuclear doctrine. It also sought to elevate the strategic positioning of Russian forces in Kursk and Crimea. The campaign extended its reach by exploiting global and regional tensions. It attempted to create discord among Western allies, particularly by leveraging the Israel-Gaza conflict, and sought to deepen divisions within the European Union. In France and Germany, Operation Undercut accounts amplified far-right party rhetoric, while in broader EU discourse, the campaign fuelled resentment toward Ukraine and Brussels. Additionally, it exploited global conflicts by tying them to the 2024 U.S. elections, stoking fears of political violence and assassination attempts within the United States. The 2024 Paris Olympics were also a target, with narratives aimed at undermining confidence in the event and its organizers.

5 Applying the Framework

5.1 The Doppelgänger campaign

This section applies the attribution framework to illustrate how an attribution process can be structured and substantiated using the evidence presented in the September 2024 U.S. Department of Justice affidavit, *Affidavit in Support of the Doppelgänger Campaign Investigation.*[77]

On September 9, 2024, the FBI announced the seizure of 32 internet domains used in the Russian government-coordinated influence operation known as Doppelgänger. Since 2022, this operation has been driven by key Russian entities, including the Social Design Agency (SDA), Structura, and ANO Dialog, all operating under the oversight of the Russian Presidential Administration, led by Sergey Kiriyenko. These entities orchestrated highly coordinated influence operations designed to erode international support for Ukraine, amplify pro-Russian narratives, and to influence voter perceptions in the United States and other nations. Doppelgänger operated by impersonating legitimate news outlets and private individuals, disseminating Russian government propaganda under the appearance of independent journalism. By mimicking trusted sources, the operation aimed to manipulate public opinion, undermine trust in democratic institutions, and to bolster Russia´s geopolitical agenda. The activities exemplify the complex interplay of infrastructure, strategy, and messaging in contemporary influence operations, making it a critical case study for attribution efforts.

The Doppelgänger campaign consisted of two interconnected operations, each carefully designed to influence public opinion while obscuring ties to the Russian government. The first involved the creation of fake websites that replicate major media outlets, such as *The Washington Post* and *Fox News*. These cybersquatted domains were crafted to closely mirror legitimate websites in appearance, including the use of authentic-looking design elements, logos, and even bylines. However, the core content consisted of fabricated articles aligned with Russian strategic narratives. While some links on these sites redirected users to genuine outlets, most of them featured misleading or false information aimed at shaping public opinion, particularly in the United States and Europe, while concealing the operation's Kremlin origins.[78] The second operation focused on amplifying this fabricated content through coordinated social media activity. Doppelgänger employed fake profiles that impersonated non-

[77] U.S. Department of Justice. (2024). *Affidavit in Support of the Doppelgänger Campaign Investigation*.

[78] Ibid., p. 14–15

Russian citizens to disseminate links to the counterfeit websites. These links were embedded in posts designed to appear authentic, furthering the campaign's aim of sowing disinformation and confusion on a wider scale.[79]

5.1.1 Technical Evidence

The technical evidence in the affidavit lays out direct operational links between Doppelgänger and sanctioned Russian entities, exposing the infrastructure behind the campaign.

Infrastructure and Domain Ownership: Doppelgänger utilized more than 60 cybersquatted domains, including washingtonpost[.]pm and foxnews[.] that closely mimicked legitimate news websites.[80] The domains resembled design elements, logos, and journalist bylines, producing fake versions of trusted outlets that were almost impossible to tell apart. Some links on these sites led to genuine sites, cloaking the deceptive origins of the campaign even further. According to the affidavit, individuals linked to Doppelgänger, working under the auspices of the Russian Presidential Administration, rented these domains via registrars based in the United States. So far, the evidence found suggested that the domains were paid for from outside the U.S., implicating sanctioned individuals including Sergei Kiriyenko, Ilya Gambashidze, and Nikolai Tupikin, along with their companies, Social Design Agency (SDA) and Structura.[81] The sanctions were imposed after these actors did not secure the requisite OFAC licenses, thus violating the International Emergency Economic Powers Act.

Attributions by OFAC: On March 20, 2024, the Office of Foreign Assets Control (OFAC) sanctioned Gambashidze, Tupikin, and their companies, attributing these individuals to the roles they played in creating and operating more than 60 spoofed websites. They created sites that pretended to be legitimate news outlets, inserting false narratives into credible-looking shells. The OFAC characterized these actions as part of a long-standing malign influence campaign by the Russian Presidential Administration.[82]

Public Reporting on Doppelganger: In July 2023, the European Union (EU) sanctioned seven Russian individuals and five entities for their involvement in the Doppelgänger campaign. Among the sanctioned entities were the Social Design Agency (SDA), Structura, Gambashidze, and ANO Dialog. The EU described the campaign, known as "Recent Reliable News" (RRN), as a Russian-led digital information manipulation effort designed to disseminate propaganda supporting Russia's war against Ukraine. The operation relied on fake websites impersonating national media outlets and government organizations, as well as fake social media

[79] Ibid., p. 16–17

[80] Ibid., p. 14–15

[81] Ibid., p. 12–13

[82] Ibid., p. 12

accounts to amplify its reach. Structura and SDA were identified as key players in creating these fake websites and boosting the campaign on social media. France's Viginum Agency, responsible for monitoring foreign digital interference, further emphasized Doppelgänger's tactics, including the creation of typosquatted[83] domains that closely mimicked legitimate media websites. These sites used the same source code as authentic outlets which made them appear credible to unsuspecting users. Since February 2023, Viginum reported over 160 Facebook pages linked to the campaign, publishing more than 600 sponsored posts containing links to fake articles and websites. Doppelgänger also used social media advertisements targeting U.S. politicians and leveraged artificial intelligence to generate disinformation content, underscoring the campaign's sophisticated and wide-reaching methods.[84]

Evidence of direct involvement and oversight by the Russian Presidential Administration: The investigation into the Doppelgänger campaign revealed extensive evidence of direct involvement and oversight by the Russian Presidential Administration, particularly through Sergei Vladilenovich Kiriyenko, often referred to as "Putin's right-hand man." Notes taken by Ilya Gambashidze, a key figure in the operation, documented at least 20 meetings between April 2022 and April 2023 that involved Kiriyenko, the Social Design Agency (SDA), Structura, ANO Dialog, and other participants. These meetings detailed strategic planning for influence operations targeting foreign and domestic audiences. For example, in an April 16, 2022 meeting titled "Meeting with SVK at the AP," Kiriyenko, referred to as "SVK" in the notes, emphasized using exaggerated narratives, including the creation of a "nuclear psychosis" to influence Western perceptions of Russia's invasion of Ukraine. Another meeting in July 2022 revealed plans to focus its propaganda efforts on Germany, with Kiriyenko stressing the need to discredit the U.S., the U.K., and NATO while convincing Germans to oppose sanctions on Russia. Specific tactics discussed included creating fake news narratives, such as fabricating stories about American soldiers committing crimes in Germany. Doppelgänger's activities were not limited to foreign influence operations as evidence suggests that it also engaged in domestic influence efforts within Russia, further highlighting its deep connection to the Russian government. For instance, one note explicitly stated that the "project could be used for P's election campaign," which, according to the affidavit, is likely a reference to Russian President Vladimir Putin.[85]

Additionally, an internal SDA document titled "Countermeasures by Foreign Agencies and Organizations" highlighted the growing concern among Western countries regarding the effectiveness of the Doppelgänger campaign. The document noted that since September 2022, the "collective West," along with major online platforms, fact-checkers, and investigators, had actively worked to counter the campaign's narratives.

[83] Typosquatting is the act of registering a common misspelling of another organization's domain as your own.

[84] Ibid., p. 19-20

[85] Ibid, p. 20-24

It also detailed 15 publications from various media outlets and organizations, including Meta, Premier Ministre, The Washington Post, Wired, and Le Monde, that reported on Doppelgänger. The investigator believed this reflects SDAs acknowledgment of its role in Doppelganger. The SDA documents was lawfully obtained during the investigation.[86]

Russian links to the leased cybersquatted domains: The FBI's investigation uncovered that Doppelgänger leased numerous cybersquatted domains from U.S. companies, including Namecheap, NameSilo, and GoDaddy, using four distinct online personas. These personas utilized email accounts incorporating their respective names and exhibited significant overlap in the legitimate news outlets their cybersquatted domains impersonated. All four personas shared notable patterns, including cryptocurrency payments and the use of Proton Mail email addresses. Cryptocurrency analysis tools and expertise from an FBI cryptocurrency subject matter expert revealed that transactions to the domain registrars originated from a cluster of cryptocurrency wallets. This cluster was funded through an account at a virtual currency exchange (VCE-1) linked to one individual. Records showed that this individual had submitted Russian identification documents to VCE-1 and accessed his account exclusively from Russian IP addresses. When interviewed by U.S. law enforcement, he described his role as a "point-to-point exchanger," claiming he had no knowledge of the origin of the funds he handled. This investigation established probable cause to believe that the funds used to lease these domains originated outside the United States.

Login activity for the registrar accounts tied to these personas corresponded with Moscow business hours, and all IP addresses used to access the registrars resolved to virtual private server (VPS) services or compromised IP addresses previously associated with cybercriminal activity. The personas demonstrated significant efforts to obscure their identities, including layering multiple VPS services and using operational email addresses with fake identifying information. For example, the Kamcopec persona utilized at least three layers of VPS services, a tactic indicative of high technical sophistication and deliberate efforts to conceal identities.

The investigation also revealed that Doppelgänger's activities were consistent with instructions in internal SDA documents, which emphasized minimizing the detection of a "Russian footprint" by employing a multi-level security infrastructure, including VPNs and U.S.-based servers. The tactics used to lease and operate these domains, combined with evidence of coordination, made the FBI suggest that the personas were acting on behalf of sanctioned entities such as SDA, STRUCTURA, and ANO Dialog and that these actions were likely carried out under the direction of Sergey Kiriyenko, a sanctioned Russian official, and the Russian government.[87]

[86] Ibid., p. 25

[87] Ibid., p. 37

5.1.2 Behavioural Evidence

The behavioural evidence focuses on the tactics and techniques that was employed by Doppelgänger to distribute its narratives but also to conceal its origins.

Coordinated social media activity: The FBI used the WayBack Machine, a digital archive, to locate and reviews articles published on Doppelgänger´s cybersquatted domains such as washintonpost[.]pm, which is a nearly identical duplication of the legitimate Washinton Post website. The duplication is believed to be created to mislead or confuse persons in the US into believing that the false or misleading information presented is from a legitimate US-based news source.[88] Doppelgänger employed deceptive tactics to distribute its propaganda while obscuring its ties to the Russian government.[89] The SDA documents outlined a comprehensive strategy for executing the Doppelgänger campaign, which relied heavily on impersonation and deception. A core tactic involved creating fake social media profiles posing as U.S. or other non-Russian citizens. These accounts were used to distribute pro-Russian narratives by posting comments or sharing content that linked to cybersquatted domains mimicking legitimate news outlets like *The Washington Post* or *Fox News*. This approach aimed to mislead audiences, particularly in the U.S., into believing they were engaging with the perspectives of fellow citizens rather than Russian state-sponsored propaganda. The documents also included specific instructions for managing these fake profiles, complete with exemplar posts and detailed scripts. For instance, one plan envisioned articles published on cybersquatted domains with headlines like "U.S. Loses Its Position as a World Leader by Making Too Many Mistakes," which would be distributed by profiles claiming to be Americans from small towns. These fake accounts were designed to appear relatable and credible to influence American voters, particularly around election periods.[90]

The links created for the cybersquat domains were crafted to appear legitimate, directing users to websites mimicking credible news outlets, such as *washingtonpost[.]pm* or *foxnews.cx*. While visiting the base domain (e.g., *www.washingtonpost[.]pm*) would result in a blank or error page, article-specific links led to active pages hosting disinformation alongside links that re-routed to legitimate outlets to reinforce credibility. To amplify its reach, Doppelgänger also purchased advertisements on social media platforms, driving traffic to these fabricated articles. This approach aimed to conceal from targeted readers, particularly Americans, that they were not accessing legitimate news sources. Additionally, the campaign extended to original media brands created by ANO Dialog and TABAK, under the direction of Sergey Kiriyenko and the Russian government. These entities developed seemingly independent journalist personas or news organizations that were publishing Russian propaganda. The same articles frequently appeared across both the cybersquatted

[88] Ibid., p. 15–16

[89] Ibid., p. 16–17

[90] Ibid., 26–27

domains, and the ANO Dialog brands, signalling close coordination between ANO Dialog, SDA, and Structura, all operating under Kremlin oversight.[91]

According to the SDA documents, particular campaigns had different plans for achieving its goals. For the campaign targeting the 2024 US elections, for instance, referred to as "The Good Old U.S.A Project", the content distributed by SDA was explicitly described as "bogus stories disguised as newsworthy events," designed to appear credible while promoting false narratives. The plan also included a "commentary campaign," involving the mass distribution of text comments and memes on platforms like Facebook and X (formerly Twitter).[92]

Strategies to circumvent mitigation efforts: One of the obtained SDA documents revealed that social media companies had attempted to counter the propaganda efforts of SDA, Structura, and ANO Dialog by flagging and blocking cybersquatted domains, including those linked to the Recent Reliable News (RRN) campaign. In response, the SDA devised a strategy to bypass these disruptions by combining automated bots with human-operated social media accounts to spread their narratives. The plan aimed to generate 60,000 comments monthly across platforms targeting audiences in France and Germany, further amplifying their messaging and undermining social media platforms mitigation efforts.[93]

5.1.3 Contextual Evidence

The contextual evidence underscores the content and narratives that has been disseminated by Doppelgänger which reveals a clear alignment with Kremlin objectives.

Narrative themes: The contextual evidence primarily focuses on the narratives disseminated by Doppelgänger. Articles published on its fake websites consistently promoted pro-Russian and anti-Ukrainian messaging. Examples include discrediting Ukrainian leadership, questioning the effectiveness of Western military aid, and advancing conspiracy theories about U.S. election interference. These narratives align closely with the Kremlin's strategic objectives. For example, the content published on the cybersquatted domain *washingtonpost[.]pm* reflects a clear pro-Russia and anti-Ukrainian stance, with many articles focusing on U.S. policy or politics. These articles do not include explicit attribution to the Social Design Agency (SDA), Structura, or the Russian government. For example, one article, titled *"White House Miscalculated: Conflict with Ukraine Strengthens Russia,"* falsely claims to be authored by a Washington Post journalist. It criticizes U.S. support for Ukraine, describing it as a waste of lives and money, and urges the Biden administration to abandon its support and pursue a peace agreement. This fabricated narrative aligns with broader Russian

[91] Ibid., p. 16–17

[92] Ibid., p. 30

[93] Ibid., 25

propaganda efforts to undermine U.S. policies and weaken international support for Ukraine.[94]

By targeting politically sensitive issues, such as U.S. elections and EU solidarity, Doppelgänger aimed to polarize societies and undermine trust in democratic institutions. The fake news articles produced by the campaign frequently echoed those narratives spread by RT and Sputnik, thus reinforcing its inclusion in the larger Russian disinformation system.[95] Moreover, contextual evidence underscores how the Doppelgänger campaign tailored its approach to suit specific target audiences. For example, in the "Meeting Minutes AP_25.07.22 – 11.00," Sergey Kiriyenko and Kremlin officials prioritized shaping German public opinion by framing Western allies as the root cause of strained relations.[96] Another note, titled "Meeting Minutes_-_AP_Factory_01.27.23," emphasized the importance of tailoring narratives to appear as though they originated from within the target audience, stating that, "When providing a narrative, it's important to remember that this is 'from a German to a German,' 'from a Frenchman to a Frenchman.'" According to the affidavit, this reflects the Doppelgänger campaign's strategy of impersonating citizens of other countries to more effectively sway public opinion and influence target audiences.[97] In one document, SDA detailed its campaign targeting Germany, organizing its efforts around three key themes: "HOHLI – pigs," "The difference between Ukraine and Germany," and "The U.S. is behind everything." The campaign focused on fostering hostility toward Ukraine by emphasizing the cultural and political differences between Ukraine and Germany as well as promoting anti-American sentiment. The document included 43 specific propaganda ideas, each associated with one or more of these overarching themes. These ideas were systematically organized into a table, which outlined the intended target audiences and the types of media to be used for dissemination.[98]

Other examples include an SDA document titled *International Conflict Incitement* outlined a campaign targeting Germany and France with the goal of escalating internal tensions in nations allied with the United States[99] or how the "The Good Old U.S.A. Project" aimed to shift U.S. public opinion toward prioritizing domestic issues over international engagements, such as financial and military support for Ukraine. The project outlined objectives and identified specific demographics to target, focusing particularly on voters in six key swing states.[100]

[94] Ibid., p. 16

[95] Ibid., p. 15–16

[96] Ibid., p. 23

[97] Ibid., p. 24–25

[98] Ibid., p. 29

[99] Ibid., p. 29

[100] Ibid., p. 30

5.1.4 Conclusion

The Doppelgänger campaign exemplifies a highly coordinated Russian influence operation leveraging fake websites and deceptive social media tactics to manipulate public perception. By mimicking legitimate media outlets and embedding fabricated narratives into seemingly credible news platforms, the operation sought to erode trust in democratic institutions and amplify pro-Kremlin messaging. The attribution of the campaign to the Social Design Agency (SDA) and associated Russian actors was supported by technical, behavioural, and contextual evidence

To illustrate the application of the IIO Attribution Framework, the evidence from the Doppelgänger campaign has been systematically mapped using the framework. The evidence presented has been categorized within the framework below and demonstrates how different sources and types of information can contribute to attribution assessments.

	Technical evidence	Behavioural evidence	Contextual evidence	Legal & ethical assessment
Open source	Domain registrations of cybersquatted websites; hosting data linked to Russian IPs and Moscow business hours; cryptocurrency wallets associated with domains.	Public dissemination of fabricated articles on social media platforms using fake accounts; automated bot activity boosting fake narratives.	Narratives supporting Russian objectives, such as discrediting Western aid to Ukraine or amplifying anti-Ukrainian sentiment; publicly visible connections to broader Russian disinformation efforts.	Ethical concerns regarding the publication of personal data and risks of misattribution.
Proprietary source	Metadata linking hosting servers and domains to sanctioned Russian entities like SDA and Structura; OFAC sanctions linking individuals to fake websites; registrar logs showing consistent access from Russian IPs; evidence of cryptocurrency payments traced via blockchain analysis.	Patterns of coordinated posting; amplification of fake news articles through paid advertisements; evidence of Doppelgänger accounts promoting narratives on multiple platforms.	Coordination between cybersquatted domains and Russian state media narratives; records of engagement between Russian-backed entities.	Challenges in handling platform data while ensuring compliance with privacy regulations; risks of misinterpretation in attribution efforts.
Classified source	Tracing of domain operations to Kremlin-linked entities; records of direct communications between sanctioned individuals such as Sergei Kiriyenko and SDA operatives.	Detailed scripts and posting schedules uncovered in internal SDA documents; evidence of coordination between fake websites and real social media amplification; patterns of ads targeting U.S. and European audiences with misleading links.	Internal Russian documents exposing projects such as the "Good Old U.S.A Project" and tailored campaigns against Germany and France; alignment with Kremlin objectives to polarize Western societies and undermine democratic institutions.	The handling of classified information requires adherence to national security laws and careful sharing with stakeholders to prevent unintended risks.

Figure 5: IIO Attribution Framework applied to the Doppelgänger campaign

5.2 The LVU campaign

In the report *LVU-kampanjen: Desinformation, kospirationsteorier, och kopplingarna mellan det inhemska och det internationella i relation till informationspåverkan från icke-statliga aktörer*,[101] Magnus Ranstorp and Linda Ahlerup presents a chronological account of the emergence and development of the so-called LVU campaign, both in Sweden and internationally, as well as the connections between these dimensions. The LVU campaign was a coordinated disinformation effort targeting Sweden, specifically its social services and the legal framework surrounding the care of minors (LVU). It falsely accused Swedish authorities of kidnapping children, particularly those of foreign descent or Muslim faith. The campaign originated from domestic actors and expanded internationally through social media and physical demonstrations, leveraging connections with radical Islamist networks and foreign influencers.

The campaign escalated into what was described as the largest information influence campaign towards Sweden,[102] spreading conspiracy theories and inciting hate against social workers and institutions. It also provoked threats of violence and terror attacks. Narratives ranged from claims of child abuse in custody to allegations of cultural erasure which exploited societal mistrust in public institutions and legal systems. The campaign combined local and global elements and blended disinformation with ideological and geopolitical agendas.

5.2.1 Technical Evidence

The LVU campaign used Arabic- and Turkish-language social media platforms, such as Facebook, YouTube, and WhatsApp, to propagate its central narrative that the Swedish authorities were systematically "stealing" Muslim children. The content included emotionally charged videos, false testimonials, and salacious graphics charging systemic Islamophobia in Sweden. Analysis of platform data made by the authors revealed patterns of inauthentic behaviour such as the repeated postings from newly created accounts and the use of cloaking techniques that sought to evade moderation. Shuoun Islamiya, an online forum commonly used to discuss Islamic topics, was one of the major hubs of disinformation amplification.[103] Moreover, the campaign featured the use of professionally produced videos which indicated that the actors had access to advanced editing tools. It also suggested coordination with well-

[101] Ranstorp, M. & Ahlerup, L. (2023). LVU-kampanjen: Desinformation, konspirationsteorier, och kopplingarna mellan det inhemska och det internationella i relation till informationspåverkan från icke-statliga aktörer. Swedish Defence University.

[102] Government Offices of Sweden. (2023). *Government Taking Strong Action Against Disinformation and Rumour-Spreading Campaign*. Retrieved from https://www.government.se/press-releases/2023/02/government-takingstrong-action-against-disinformation-and-rumour-spreading-campaign/

[103] Ranstorp, M. & Ahlerup, L. (2023) *LVU-kampanjen: Desinformation, konspirationsteorier, och kopplingarna mellan det inhemska och det internationella i relation till informationspåverkan från icke-statliga aktörer*, p. 12–13. Swedish Defence University.

resourced actors.[104] The campaign also used international Arabic-language media, including outlets with acknowledged Islamist orientations to spread its message and articles and opinion pieces reflected disinformation narratives which gave the claims additional credibility at the same time.[105]

5.2.2 Behavioural Evidence

To unify the campaign message, conductive user logs were analysed to use different platforms to reach a wide variety of people. Videos and posts were first published on platforms such as YouTube and Facebook and then re-emitted for specific communities via WhatsApp and Telegram.[106] One example of this cross-platform dissemination and coordinated timing was the quick way a video went viral featuring a crying mother asking for the return of her children. Within hours, the video travelled through multiple groups on Facebook and channels on WhatsApp, with the same captions in Arabic, Turkish, and English. Moreover, certain posts used cloaking to evade content moderation, sending users to other landing pages depending on their location and device type.[107]

Another defining feature of the campaign was its reliance on amplifying content through networks of social media accounts. This was discovered through an analysis of shared posts linked to LVU-related content which revealed that over 100 accounts actively engaged in disseminating these messages and that they often used coordinated timing, similar language, or consistent hashtags. This suggests the use of cross-posting tools[108]. A notable example involved the systematic management of a cluster of accounts overseen by administrators in Middle Eastern countries. These accounts consistently posted about alleged abductions and frequently tagged prominent Islamic scholars and organizations to increase visibility and engagement.[109]

The campaign had some prominent actors. For instance, Moustafa El-Sharqawy played a central role in the campaign and used his platform to depict the LVU as a calculated assault on Islam. His efforts included the distributing of videos, organizing protests, and partnering with other influencers to magnify the campaign's message.[110] Furthermore, religious leaders across various countries joined the movement, encouraging their congregations to denounce Sweden's policies. In some cases, Friday

[104] Ibid., 15–16

[105] Ibid., p. 22

[106] Ibid., p. 18

[107] Ibid., p. 20

[108] Ibid., p. 23–25

[109] Ibid., p. 27

[110] Ibid., p. 29–30

sermons in mosques featured LVU-related accusations and intertwined the campaign with narratives that is commonly associated with radical Islamist rhetoric.[111]

The campaign also extended into physical demonstrations, including protests held outside Swedish embassies in Iraq and Turkey. These events were livestreamed and quickly transforming into viral online content.[112]

5.2.3 Contextual Evidence

The central narrative of the LVU campaign portrayed the Swedish state as systematically hostile toward Muslim families and the campaign capitalized on fears of cultural assimilation and Islamophobia. Unverified stories of children allegedly being forced to eat pork or abandon Islamic traditions circulated widely on social media, fuelling outrage and amplifying the campaign's reach.[113] Extremist groups leveraged the LVU campaign to reinforce their anti-Western rhetoric and organizations tied to radical ideologies, such as Shuoun Islamiya, used the campaign to propagate broader narratives about the persecution of Muslims in Europe.[114] For instance, a radical preacher released a video urging Muslims to view Sweden as part of a "global conspiracy against Islam," integrating the LVU claims into a larger anti-Western discourse.[115]

The campaign's impact extended far beyond Sweden's borders. Turkish media, for example, amplified the LVU allegations, often citing Arabic sources, while hashtags like #SaveMuslimChildren gained traction across multiple languages. This transnational spread created a feedback loop, where the narrative was continually reinforced through repetition across social media platforms and international publications which increased its legitimacy.[116]

5.2.4 Conclusion

The campaign's involvement of multiple actors, ranging from domestic activists, international influencers, to extremist networks, pose a challenge to directly link specific entities to the campaign. However, combining technical (e.g., IP analysis), behavioural (e.g., amplification patterns), and contextual (e.g., narrative resonance) evidence provided a comprehensive attribution case for the LVU campaign as a coordinated campaign and not a collection of unrelated or spontaneous activities. This

[111] Ibid., p. 32

[112] Ibid., p. 34

[113] Ibid., p. 35–37

[114] Ibid., p. 38–40

[115] Ibid., p. 41

[116] Ibid., p. 42–43

alignment of evidence demonstrates the strategic intent and organized nature of the campaign.

To illustrate the application of the IIO Attribution Framework, the evidence from the LVU campaign has been systematically mapped using the framework. The evidence presented has been categorized within the framework below and demonstrates how different sources and types of information can contribute to attribution assessments.

	Technical evidence	Behavioural evidence	Contextual evidence	Legal & ethical assessment
Open source	Use of Arabic- and Turkish-language social media platforms (Facebook, YouTube, WhatsApp); emotionally charged videos and testimonials; cloaking techniques to evade moderation.	Rapid dissemination of viral content across multiple platforms; network analysis revealing coordinated amplification by over 100 accounts; patterns of consistent hashtags and cross-posting strategies visible on public platforms.	Narratives framing Sweden as hostile to Muslim families; unverified allegations of cultural assimilation propagated via viral posts and hashtags (#SaveMuslimChildren); visible ties to extremist rhetoric in public social media content.	Ethical concerns about analyzing and publishing data that involve personal details; risk of harm to individuals conducting analyses or exposing such campaigns. .
Proprietary source	Not used	Not used	Not used	Not used
Classified source	Not used	Not used	Not used	Not used

Figure 6: IIO Attribution Framework applied to the LVU campaign

5.3 The Paperwall campaign

Paperwall was a sophisticated influence operation attributed to Shenzhen Haimaiyunxiang Media Co., Ltd. (Haimai) by The Citizen Lab at the University of Toronto[117], and can be systematically categorized within the IIO Attribution Framework structure. The campaign leveraged over 123 websites impersonating local news outlets in 30 countries to disseminate pro-Beijing propaganda.

5.3.1 Technical Evidence

The technical evidence was a key component of the attribution process. The Citizen Lab traced the campaign's digital infrastructure directly to Shenzhen Haimaiyunxiang Media Co., Ltd., with domain registrations and server hosting linked to Haimai's operational network. WHOIS data revealed overlapping registrants between Haimai and the campaign websites which demonstrated centralized management. Additionally, the operation´s websites mimicked legitimate local news outlets with remarkable accuracy. It employed professional layouts, logos, and localized branding and the source codes of these websites often mirrored those of authentic outlets, showcasing a deliberate effort to deceive users into trusting them. Content analysis further revealed that many articles published on these websites originated from Times Newswire, a distribution service that has been implicated in previous Chinese influence campaigns. Metadata embedded in the content confirmed its connection to Times Newswire servers, similar to what had been identified in earlier state-aligned efforts. Hosting servers for the operation were predominantly located in China, with IP addresses resolving to providers that had previously been associated with state-sponsored activities and the registrars used were consistent with those in past influence campaigns, reinforcing the technical links to Chinese actors.

5.3.2 Behavioural Evidence

Behavioural evidence provided additional insights into the operation´s coordination and tactics. Articles and narratives were disseminated across the network in a synchronized manner with simultaneous postings on multiple websites which suggested centralized scheduling and operational control. Timestamps aligned with Chinese business hours which increased the suspicions of a coordinated effort originating in China. A notable behavioural tactic involved the strategic removal of aggressive content, such as ad hominem attacks on Beijing's critics, shortly after publication. The researchers at The Citizen Lab recovered cached versions of deleted material and could expose a deliberate strategy to maximize initial impact while still evading detection.

[117] The Citizen Lab. (2024). *PAPERWALL: Chinese Websites Posing as Local News Outlets with Pro-Beijing Content*. University of Toronto.

The operation also made use of localized adaptation by tailoring its messaging to resonate with specific regional audiences. For example, in Latin America, articles emphasized China's economic investments, while in Europe, content discredited criticisms of China's human rights record. This adaptability highlighted the operation´s sophisticated understanding of audience segmentation. Additionally, the integration of benign press releases with propaganda allowed the operation to blend into the broader media ecosystem, creating plausible deniability and complicating immediate identification as an influence operation.

5.3.3 Contextual Evidence

Contextual evidence further supported the attribution. The narratives promoted by the operation consistently aligned with Beijing's strategic goals, including promoting the Belt and Road Initiative as a benevolent global investment program, downplaying international concerns about human rights abuses in Xinjiang and Hong Kong, and amplifying anti-Western sentiment by framing the U.S. and Europe as imperialist and hypocritical. The campaign's tactics closely mirrored those observed in previous Chinese influence operations, such as the use of Times Newswire and similar dissemination networks, underscoring a recurring pattern of state-aligned activities. By impersonating local news outlets, the campaign seamlessly integrated into regional media ecosystems, increasing its credibility among target audiences while obscuring its Chinese origins. The direct involvement of Haimai, a private PR firm, highlighted China's strategy of outsourcing influence operations to commercial entities. Financial records tied Haimai to hosting services for the campaign domains, showcasing the firm's central role in the operation.

5.3.4 Conclusion

The Paperwall campaign represents a calculated effort by Shenzhen Haimaiyunxiang Media Co., Ltd. (Haimai) to embed pro-Beijing narratives into local media ecosystems worldwide. Through the creation of over 123 fake news websites across 30 countries, the operation effectively disguised state-aligned messaging as independent journalism. Technical evidence, such as domain registrations and hosting patterns, linked the campaign to Chinese interests, while behavioral analysis revealed synchronized content dissemination strategies tailored to regional audiences. Contextual evidence shows that the campaign's narratives aligned closely with Beijing's strategic goals

To illustrate the application of the IIO Attribution Framework, the evidence from the Paperwall campaign has been systematically mapped using the framework. The evidence presented has been categorized within the framework below and demonstrates how different sources and types of information can contribute to attribution assessments.

	Technical evidence	Behavioural evidence	Contextual evidence	Legal & ethical assessment
Open source	Domain registrations linked to Shenzhen Haimaiyunxiang Media Co., Ltd.; metadata connecting content to Times Newswire servers; hosting servers and IP addresses resolving to China.	Articles posted in synchronization across 123 websites; timestamps aligning with Chinese business hours; localized messaging for specific regions, such as Latin America or Europe to resonate with local sentiments; integration of benign press releases with state-aligned narratives, blending propaganda into legitimate media ecosystems.	Narratives consistently aligned with Beijing's strategic goals, including promoting the Belt and Road Initiative, downplaying human rights abuses, and amplifying anti-Western sentiment; tactics mirroring previous Chinese information influence operations.	Risk of misattributing state involvement when private actors like Haimai are involved.
Proprietary source	WHOIS data and metadata linking Shenzhen Haimaiyunxiang Media Co., Ltd. to campaign domains; platform analytics revealing centralized scheduling and operational control; evidence of coordinated use of registrars previously linked to Chinese influence operations; financial records tying Haimai to hosting services for campaign domains.	Strategic removal of content, such as ad hominem attacks, shortly after publication.	Not used	Legal aspects of collecting data from server providers to verify connections to actors involved
Classified source	Not used	Not used	Not used	Not used

Figure 7: IIO Attribution Framework applied to the Paperwall campaign

6 Critical discussion

The NATO Stratcom CoE and Hybrid CoE IIO Attribution Framework works. It has been tested both by members of the counter-FIMI community, and in this report in the form of case studies. We do not identify a need for significant modifications, even though certain considerations have warranted further reflection resulting in changes or additions.

A key observation is that most organisations prefer to remove the section on classified evidence sources due to them not having access to intelligence (or not wanting to disclose that access). In the majority of cases that we have seen, investigating organisations use OSINT rather than declassifying intelligence. However, on balance we feel that it is useful to retain these parts of the matrix, first so that the framework includes organisations that work with secret intelligence, and second to remind those who use it of the bigger picture.

The second small amendment that we think is necessary is to draw a clearer separation between the first three columns and the fourth. We feel increasingly that legal and ethical considerations deserve a place in the matrix, but that they are qualitatively different activities, often conducted by different parts of an organisation. We demonstrate this distinction by drawing a thicker line between this column and the others.

A third small change is the introduction of a traffic light colour-coding scheme for the evidence categorization, which we believe further increases the flexibility of the framework as a communicative tool: an organisation could for example only use the colours to give a brief overview of what types of evidence they have considered. Finally, we have explicitly mentioned financial data as a sub-category to technical evidence, on the basis that it is a key part of investigative work.

One of the emerging trends that we have observed, especially since the Russian invasion of Ukraine, is the increasing importance of attribution as a societal function. Prior to the invasion, the release of declassified intelligence about Russian efforts to justify the war (so-called prebunking) were a first example of this. The widespread sanctions against Russian state media were a second. The continued exposure of Doppelgänger and associated campaigns a third. Put simply, attribution is becoming more important to the counter-FIMI toolbox than ever before. We therefore advocate for the more widespread use of a framework such as this to simplify the understanding of attributions for both the counter-FIMI community as well as the general public.

Of course, any attribution risks revealing the methods by which information was obtained. This is known as tradecraft. In many respects, a tool such as the attribution framework gives the attributing organization power over the information they reveal, while maximizing transparency. It also opens for interoperability. The more that the counter-FIMI community shares tools, and in particular makes use of structured methods, the more a disparate, eclectic group of actors will behave as a community. We are all part of a networked "defender community" rather than single organizations able to understand and do everything on their own. In order for such a community to work, we must commit to some basic standards. In our view, the NATO Stratcom CoE and Hybrid CoE IIO Attribution Framework represents a strong foundation for public attribution efforts.