978-9934-619-14-4

# ATTRIBUTING INFORMATION INFLUENCE OPERATIONS

## IDENTIFYING THOSE RESPONSIBLE FOR MALICIOUS BEHAVIOUR ONLINE

Hybrid CoE


A joint report from the NATO Strategic Communications Centre of Excellence and the European Centre of Excellence for Countering Hybrid Threats.

ISBN: 978-9934-619-14-4
Authors: James Pamment, Victoria Smith
Project Manager: Ben Heap
Design: Kārlis Ulmanis

Riga, July 2022
NATO STRATCOM COE
11b Kalnciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook/stratcomcoe
Twitter: @stratcomcoe

# Foreword

By their nature hybrid threats create ambiguity. They are covert, obscuring intent and enabling plausible deniability. This makes identifying both who is behind them and their motives a significant challenge.

Attribution must grapple with this challenge. It involves identifying the responsible actors, understanding what they are hoping to achieve and how they are seeking to accomplish it. Without a methodical and consistent approach underpinning attribution, planning and executing effective responses is far more challenging.

The online environment has become a rich source of opportunities for a type of hybrid threat, referred to in this report as 'Information Influence Operations'. These are deliberate attempts at deception, including interference in democratic processes, using disruptive and illegitimate means which are readily available to hostile actors.

Those responsible for malicious behaviour online are motivated by a wide range of reasons. Organised, state-sponsored instruments of influence work across borders, furthering strategic aims by reaching directly into a targeted nation's society and political structures. Others act for simpler reasons such as excitement, financial gain or self-actualisation. Working out who's who means untangling a variety of evidence and indicators.

Threats which occur in the digital domain pose a particular problem. By their very nature, online platforms are susceptible to manipulation, and even though responses such as takedowns are increasingly common, more work is needed to improve the process of identifying those responsible for malicious behaviour. A cat-and-mouse game has developed between hostile

actors, governments, researchers and technology companies. The supporting concepts have not always kept pace with emerging threats.

In this joint report from the NATO Strategic Communications Centre of Excellence and the European Centre of Excellence for Countering Hybrid Threats, James Pamment and Victoria Smith provide a solid base for further discussion on the attribution of information influence operations, as part of an overall effort between our two Centres of Excellence to improve the theory and practice of attribution. This paper aims to improve the ability of practitioners to collect and analyse evidence in support of the attribution process. It reviews existing approaches to attribution and suggests that malicious activities online can be analysed using four key lenses and three key types of evidence, with differing degrees of accessibility. The paper then looks at how this framework relates to real world cases of investigation conducted by leading practitioners in the field.

Ben Heap *NATO Strategic Communications Centre of Excellence*

Stuart Mackie *European Centre of Excellence for Countering Hybrid Threats*

## Acknowledgements

Anneli Ahonen, Sebastian Bay, Carnegie Endowment for International Peace, Ben Heap, Anna-Kaisa Hiltunen, Hybrid COE Community of Interest on Hybrid Influencing, Stuart Mackie, Björn Palmertz, Henrik Twetman, Gavin Wilde

# Contents

> Attribution is a crucial component of Information Influence Operation analysis because it is the stage at which threat actors are held to account for their actions.

# Introduction

Digital platforms are vulnerable to manipulation. Since 2018, the "big three" platforms have announced over 350 "takedowns" of coordinated efforts to manipulate their platforms.[1] Such takedowns usually involve a statement of blame toward the actors behind the platform manipulation. This is known as an attribution. At its heart, attribution is the act of determining who is responsible for specific illicit actions or outcomes. While blame can be apportioned with or without supporting evidence, attribution is used in this report to describe the methodical process by which evidence of Information Influence Operation (IIO) activity is collected, assessed, and approved for communication to the public. Further nuance in definitions will appear throughout this report, and will be addressed more fully in future working papers.

Platform manipulation takes many forms and is therefore called many different things. In this report, we prefer the term Information Influence Operation (IIO). IIO is the organised attempt to achieve a specific effect among a target audience, often using illegitimate and manipulative behaviour. They "exploit open and free opinion-formation by mimicking legitimate behaviour to gain access to and influence the public sphere."[2] Those carrying out IIO draw on communicative tactics such as fabrication, false identities, malign rhetoric, symbolism, and technological advantages to exploit vulnerabilities in the information environment.[3] Implicit in the definition of IIO is the assumption that one or more actors have planned and conducted an operation that serves the interests of, for example, a hostile state. Attribution is therefore a crucial component of IIO analysis because it is the stage at which threat actors are held to account for their actions.

Attributing IIO can at its most transparent involve presenting compelling evidence that an organisation such as a public relations company, political party, or state actor

was behind the manipulation effort; at its most opaque, it is finger-pointing without any shared supporting evidence. As this report will outline, attribution of IIO is in a dysfunctional state. A typical platform takedown makes a statement of attribution but rarely shares the evidence that led to that conclusion. A small group of "platform approved" operational research teams are informed of an impending takedown so that they can prepare reports confirming the veracity of the platform actions. They usually do this using evidence that is suggested to them by the platforms, but that is different from the data that the platforms analysed to reach their conclusions, and that omits technical evidence. Researchers outside of this small, trusted group are by-and-large unable to research the removed content or independently verify the data that led to the takedown and attribution unless they had already found and downloaded the data prior to it being removed by the platform.[4]

Other problems in the field are more conceptual in nature. There is currently a limited conceptual language that can be used to discuss how and why IIO attribution looks the way it does. This report introduces a framework that enables more accurate discussion of the components of an attribution. In doing so, it seeks to represent the sometimes-conflicting perspectives of researchers, journalists, digital platforms, and governments and to show how each group provides crucial pieces of the attribution puzzle. The report analyses the opportunities and constraints that these members of the community face when attempting to attribute IIO. Ultimately, this report aims to open a debate about how to improve the ability of IIO analysts to assess evidence, regardless of which sector they work in. The current structures supporting IIO attribution are broken – this report is a first attempt to fix them.

This report is the point of departure for future work on IIO attribution. It is divided into three parts, each of which establishes some key principles for further development and investigation. The first part begins by reviewing the historical links between IIO and cyber attributions, arguing that while there are similarities, there are also fundamental differences that require a modified approach to IIO attributions. There follows a discussion of how the different levels of access to information affects an analyst's appetite and ability to make attribution assessments. The section ends with a short analytical overview of published IIO research, with additional data available in Annex B. Together, this first section helps to show the state of the field as it stands at present.

The second section outlines a framework for IIO attribution. The framework is based around a matrix of four types of evidence (technical, behavioural, contextual, and legal/ethical) and three sources of evidence (open source, proprietary source, and classified source). By working through each of the 12 categories that emerge from this matrix, it is possible to define in

some detail the types of considerations that are possible from the perspectives of different stakeholders engaged in IIO attribution, whether they be journalists, researchers, private sector intelligence services, digital platforms, or governments. Data from existing IIO attributions is used to support the discussion. The framework provides a terminology that can help to improve understanding between actors about the different types of evidence available, what they are able and unable to demonstrate, and to use this understanding to improve information sharing within the IIO community.

The third and final section assesses some of the limitations and opportunities open to those working with the three main data sources, with an emphasis on the potential for enhanced cooperation. Annexes offer further examples of typical IIO attribution processes (Annex A), extended data from the literature review on published IIO analysis reports (Annex B), and an overview of the research report sample (Annex C).

# 1. The state of the field

## IIO attribution is influenced by cyber frameworks

Attribution debates in IIO take their inspiration from the cyber security field. When it comes to assessing a cyberattack, technical evidence performs a critical role in raising the certainty of an attribution. Much of the existing literature on attribution concerns cyberattacks, and therefore technical analysis is the indisputable focal point.[5] The main hacker groups have identifiable patterns of behaviour and are identified as "Advanced Persistent Threats" or APT. The criteria for assessing and identifying APT is fairly well-established and standardised, so that crucial information about network vulnerabilities and attack vectors can be shared. Frameworks have helped to shape best practice in the attribution of cyberattacks, such as the Q Model,[6] the ODNI Cyber Threat Framework,[7] and The Diamond Model of Intrusion Analysis.[8]

In this report, we argue that attribution of IIO is fundamentally different from attribution in the cyber field. [9] IIO is a communication problem in which behavioural and contextual evidence are most visible. Considerations such as content, user accounts, messaging and narratives, target audiences, communicative techniques, and coordination are necessary for piecing together an understanding of strategic intent. Since IIO plays out in public discourse, it can be much harder to isolate the originator from those who amplify and add to compelling narratives. Unlike network intrusion, interjection in public debate is not illegal and is fraught with blurred lines. Additionally, much of the content of IIO can directly impact upon contextual assessments, since common goals include political polarisation, undermining public discourse, and influencing decision-makers.

Analysis of an IIO can draw on:

- **Technical evidence**, consisting of the observable traces that an adversary leaves behind at the level of digital signals;
- **Behavioural evidence**, supported by knowledge of the methods by which different adversaries carry out their work (this is often termed Tools, Techniques and Procedures or TTPs);
- **Contextual evidence**, which consists of an assessment of the content of IIO, the socio-political context in which IIO takes place, and the motivations of the adversary;
- **A legal & ethical assessment** of whether assigning blame is

proportionate, and whether it sets into motion considerations relating to e.g. political or commercial fallout, treaties or litigation.

The range of tactics and platforms used by malign actors, and the motivations and objectives that drive them are many and varied. As a result, IIO attribution is a dynamic process requiring a broad range of analytical techniques, deployed to a greater or lesser extent depending on the individual circumstance. As the field develops, it becomes crucial to focus on how to do the various types of analysis responsibly and effectively, and on how to communicate results as transparently as possible both within the wider research community and to the public.

## Different players have access to different types of data

The four types of evidence referred to above are in most cases collected from three distinct information sources. The source informs not just what information is collected and how it is assessed, but also whether and to what extent it may be made public. Evidence derived from open sources can be analysed and discussed by any actor but is typically reliant on content that is visible to the public. Evidence derived from proprietary data is rich in technical and behavioural information which is only released at the discretion of the data owners. Evidence derived from classified intelligence is either shared purely as a sanitised contextual or legal assessment, or selected parts are declassified and shared. In sum, the three information sources shaping IIO attribution are:

- **Open source**, which relies on open-source information and open-source intelligence (OSINT), and access to other publicly available information. It is used by NGOs, media, and researchers, who have little or no access to proprietary information or classified intelligence. It is also widely used by intelligence agencies in addition to classified data. Attributions frequently take the form of investigative journalism, crowd sourced research, and qualitative and quantitative content analysis of open data sets, and rely on building circumstantial cases, for example by deriving intent from the tactics and narratives used. This analysis can in some cases be strengthened by linking an adversary's activities to web domains, IP addresses, and company ownership. This helps to build a dossier of technical, behavioural, and contextual evidence that can point to an adversary's responsibility for an IIO; however, the technical information required to make a strong attribution is rarely available in open source. The ethical norm of informing the public is often of greater importance than assessments of the political or commercial fallout.

> The decision to attribute can be affected by commercial and geopolitical concerns, such as access to markets and risks of retaliatory regulation

- **Proprietary source**, which is based on privileged "backend"[10] data sources such as those available to digital platforms, private intelligence companies, data brokers, and cyber security companies. The technical and behavioural data gives insight into the infrastructure serving IIO, which allows investigators to make inferences about who is capable of coordinating such an operation. Attributions commonly take the form of platform takedowns and subscription service intelligence reports, in which the actor is revealed together with examples of their activities often derived from open sources. Given that private actors such as social media platforms are usually the holders of the technical data, their legal assessments can be linked to their own terms of service in addition to the laws of the host country. However, the decision to attribute can be affected by commercial and geopolitical concerns, such as access to markets and risks of retaliatory regulation. While these actors have access to technical data that can support a strong attribution, they are limited by the scope of their own proprietary data and are reliant on discreet data sharing partnerships to make assessments about e.g., cross-platform activities.

- **Classified source**, which is based on secret information, but can also incorporate open source and proprietary information, and is conducted primarily by governments and by extension the military. Classified information is likely to answer a narrow intelligence request about a hostile actor's behaviour and is typically prepared for internal government use, for sharing within the intelligence community, or in some instances for communication to the public via ministers, parliamentary committees, or threat intelligence summaries. They draw heavily on technical evidence but are frequently combined with behavioural, contextual, and legal assessments before an attribution is made. The goal is often to attribute IIO in the context of the broader hostile

activities of the actor in question. Proof of IIO could lead to wide ranging effects such as denunciations, diplomatic measures, or strikebacks.

An analyst's visibility of, and perspective on, an IIO will depend on their objectives, priorities, resources, and access to information. Few, if any, analysts have both access to the full spectrum of potentially available data and the resources and freedom to fully explore them. Having an awareness of the strengths and weaknesses of, and gaps in specific types of access is therefore essential to, for example, identifying potential bias or assessing probability.

## Findings from the literature review

This analysis reviewed 59 reports on IIO[11] authored by 24 different organisations,[12] and the Disinfodex database.[13] Graphika, The Atlantic Council and its Digital Forensic Research Lab (DFRLab), and Stanford's Internet Observatory comprise half of all reports analysed. This reflects their prolific and unparalleled output on this subject, but also demonstrates the dominance of US-based research organisations. These three organisations also benefit from data sharing relationships with platforms such as Facebook and Twitter. This gives them advance notice of platform takedowns so that they can time the publication of their reports with the platform's takedown announcements. In these cases, the research frequently inherits the platform attribution, which the researchers try, but do not always succeed, to independently corroborate.

In total, 19 of the reports studied made an attribution, 26 cited an attribution made by others and ten supplemented an external attribution with their own evidence. Only four reports did not include an attribution. Actors located in Russia were the most frequently attributed in the research report sample (49%). The next most commonly attributed actor was Iran (12%).

Proprietary data held by the platforms, specifically Twitter (66%) and Facebook (59%), was the single most important technical and behavioural information source for making IIO attributions in the reports studied. Yet the underlying technical data was not made available to independent researchers. Most attributions are therefore inherited: they are directed by what is found in proprietary technical data but analysed publicly at a level removed, using open sources. There may be methodological concerns with the process of recreating a takedown based on proprietary data using only open sources.

Like platforms, governments are often hesitant to publish their assessments of IIO in any detail. Russian interference in the 2016 US Presidential election is a rare exception, with the evidence published as part of an FBI inquiry. Usually, governments announce an overall assessment of the intent of a hostile actor without revealing

specific details. Sometimes, heavily redacted reports give the misleading impression that only weak data was available to analysts.

The online database of platform takedowns Disinfodex yields 520 results[14] relating to takedowns from Facebook, Google, YouTube, Reddit and Twitter. About two-thirds of these takedowns include a clear attribution to an actor. However, some attributions to sensitive targets such as governments and political parties are tempered by statements such as "individuals associated with" or "employees of". While this type of language demonstrates the role political and ethical assessments play in the framing of an attribution, it is not always clear what is meant. Terminology used to distinguish between the different types of state involvement in or direction of an IIO is often poorly defined and inconsistently applied. Rather than articulating nuance, undefined euphemisms such as "Kremlin-backed" can create confusion about the nature and extent of government involvement. Attempts to define the different levels of state involvement have been made, for example in Jason Healey's "Spectrum of State Responsibility".[15]

# 2. An IIO attribution framework

The purpose of a framework for IIO attribution is twofold; first to improve understanding between actors about the benefits and weaknesses of the different types of information available to different actors, and second to use this understanding to improve information sharing within the IIO community (journalists, researchers, NGOs, companies, intergovernmental organisations, and governments). This in turn supports the overarching aim of better informing the public so they are empowered to understand the nature of the threat from IIO. Clearly, there are many structural problems associated with how IIO is currently attributed. Our proposed framework cannot solve all the issues, but it can play a role in further making the core issues transparent as well as proposing small, realistic steps toward improvement.

The matrix below shows the four kinds of evidence 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 |

As the following examples show, the framework can give order and clarity, and potentially improve information sharing, to activities that are already standard practices in the counter-IIO field.

## Technical evidence

**The available open-source technical evidence** is generally better suited to analysing individual influence *activities*

such as disinformation than coordinated *operations* such as IIO. Technical and behavioural evidence can be gleaned from open-source intelligence techniques (OSINT), for which there are multiple off-the-shelf tools and methodologies available to the analyst. However, it is far more challenging to investigate coordination using solely open sources, as opposed to having access to a backend infrastructure. For example, in one academic study of trolling on online news platforms, the only available information about user location in the platform's public API (Application Programming Interface) was based on IP addresses.[16] Given the wide availability of VPNs capable of masking a user's location, the limited information available to researchers can cause inaccurate attributions that contribute to misunderstandings about adversary techniques. In the absence of better data, the researcher is forced to weigh up whether partial evidence is better than nothing.

Open-source technical evidence can help to find the link for example between a social media account or webpage and its owner, and reveal covert relationships between accounts and organisations. This can be complemented by technical and behavioural evidence from proprietary and classified sources that have been leaked or purchased such as cell phone data, passport registration forms, and airline ticket purchases, which for example several of Bellingcat's investigations have made use of.[17] By bringing such evidence into the open domain, the technical foundations of a public attribution are significantly strengthened. However, use of leaked datasets is mostly conducted by investigative journalists rather than researchers.

- **Links to external websites** were identified and analysed in 61% of reports in our sample
- **Domain network analysis** (searching for links between domains) is used in 20% of reports
- **Website domain analysis** (searching for domain ownership) is used in 15% of the reports

Within the research report sample, analysis of state media or other kinds of pro-state media was identified in 36 percent of reports. This demonstrates an area where open-source data can play a significant role. Since there are no covert aspects to linking technical evidence to its source in the case of state media, attribution is a relatively straightforward affair, provided that the possibility of forgery can be ruled out.[18] Other attributions made by researchers using solely open-source data account for 10% of the overall attributions.

**Proprietary technical evidence** is rarely shared with the public. The evidence builds on data that are collected and analysed by the digital platform owners. However, collecting and analysing this data often has its own internal challenges that are not

Due to the proliferation of private intelligence companies entering the IIO space, there are instances of companies approaching journalists to publish write-ups of their proprietary reports in order to gain a foothold in the commercial market or for political purposes

widely understood outside the companies themselves. These challenges include compiling datasets from information stored across multiple repositories, how responsibility for analysing different types of behaviour is allocated between team members or departments, and how many staff members are allocated to do this.

Data related for example to the creation of a new account can be analysed to reveal information about who created the account, where they are based, and potentially some of their other online activities. Further cross-analysis can reveal patterns of activity that may suggest the creation of a coordinated infrastructure capable of delivering IIO. Platforms can also set traps and build deterrents into their systems, in cat-and-mouse engagements with persistent threat actors.

When shared with the public, analysis tends to be presented as a conclusion, e.g. *individuals connected to the military and based in Russia*. For example, a comprehensive Stanford report on GRU influence operations opens by stating that Facebook was responsible for the attribution of GRU.[19] The Graphika report "From Russia with Blogs" states that the authors were unable to independently verify a social media platform's attribution to "Russian military intelligence services" due to insufficient evidence.[20] Upon request, and where a legal framework exists, such data is shared with law enforcement or intelligence agencies. Due to the proliferation of private intelligence companies entering the IIO space, there are also instances of companies approaching journalists to publish write-ups of their proprietary reports in order to gain a foothold in the commercial market or for political purposes.

**Classified technical evidence** is derived from signals intelligence (SIGINT) and human intelligence (HUMINT) that has been collected by an intelligence agency. Covert access enables an intelligence agency to collect detailed technical evidence in order to prove the relationship between an adversary and the IIO they conduct. As with proprietary evidence, if the results are

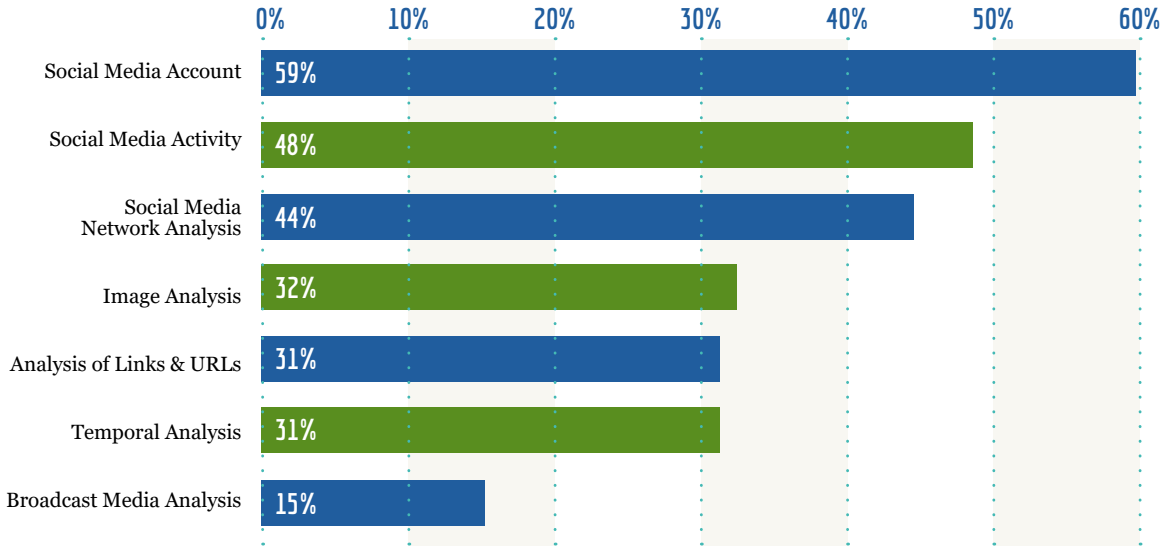


Figure 1: Most frequently identified research and analysis techniques

shared in the public domain, it is usually as a conclusion. This is to protect both personal data and the methods by which the information has been collected. For example, 40% of the 448 pages of the Mueller Report into Russian interference in the 2016 US Presidential election include some form of redaction. The section most heavily redacted was about Russian IIO, with 46% of content redacted. This was a higher level of redaction than for example the chapters on prosecution decisions and hacking.[21]

## Behavioural evidence

**Open-source behavioural evidence** tends to focus on the activities and techniques used by accounts suspected to be part of an IIO. This is a common component of research reports on IIO, covering for example account activity (including the time of day the account is active), amplification methods such as patterns of posting and cross-posting on websites, social media pages or groups, and network analysis to show connections between accounts based on behaviours such as liking, sharing, and following. Much of this data can be obtained by commercial and open-source analytical software.

**Proprietary behavioural evidence** assesses similar evidence when there is suspicion of inauthentic behaviour, with the support of technical data from the platform infrastructure. Often, this is broader than what is available through open-source methods, for example including data from closed or private accounts, groups and pages. It may also include assessment

of behaviours that are less about the IIO and more about patterns of circumventing security features implemented by platforms.

**Classified behavioural evidence** follows a similar set of practices but may be supported by intelligence identifying communications or financial trails that shed light on the connections between actors suspected of illegal activity or intent.

The analysis of research methods used in our report sample reveals that behavioural evidence is a significant source of evidence for attributions. The analysis identified the following most frequently used research and analysis techniques. This helps to reveal how IIO analyses generate data:

Techniques that research the social media account (for example when it was set up or what the profile picture and description reveal), social media activity, and image analysis (what the account is posting), social media network activity (what accounts are interacting with it and how), and temporal analysis (when the account is active), are all establishing the patterns of behaviour of the account. That five of the top seven most commonly used research techniques explore the behavioural evidence demonstrates how significant this is as a source of evidence for attributions.

The communication techniques used by an adversary can also be considered as part of a behavioural analysis. This helps to build a picture of the types of communication they are using, in order to better understand what they are doing and why.

- **Attributions to actors located in Russia, including those with connections to the Russian state,** were identified in 42 percent of reports. Communication techniques identified in our report sample include:
  - **Inauthentic amplification**[22] of content in 58%
  - **Linking to external websites** as credible sources of information in 52%
  - Efforts to **polarise** (52%) and **discredit** (31%)
  - **Trolling** in 31% of attributions citing Russian-based actors

- **Attributions connected to actors located in Iran, including those with connections to the Iranian State** comprise 10% of cases. Communication techniques identified in our report sample include:
  - General pro-Iranian **propaganda** used in 86% of attributions citing Iranian-based actors
  - **Repurposed content** and content using **memes and humour** were both identified in 71%

- **Attributions connected to actors located in China, including Chinese state actors** comprise 10% of the sample. Communication techniques identified in our report sample include:
  - Coordinated efforts to **discredit** 100%

- General pro-China **propaganda** 83%
- Use of **repurposed or hacked accounts** 67%

Analysis of the 59 papers identified 14 broad research methodologies used by the authors and 41 different tactics identified by those authors as IIO tactics. Clearly, there are many tactics and combinations of tactics used in IIO. However, it could be argued that some of the problems of attributing IIO come down to problems of research methodology (i.e. of how to reliably identify, classify, and analyse IIO behaviour) as much as problems of identification (i.e. who is responsible for the behaviour). The diffuse terminology used to describe IIO activity only makes things more confusing.

## Contextual evidence

Contextual evidence is a broad category that covers both the content of IIO and some of the political considerations that shape how that content should be understood. This builds to an assessment of the motivations and intent behind an IIO, such as electoral interference or societal polarisation. However, it is also the most difficult part of the IIO to study, since content analysis is subjective, and culturally and temporally bound. Furthermore, IIO often appeals to cognitive biases, and can be difficult to categorise consistently if it involves provocative content. Contextual evidence, sometimes in conjunction with behavioural evidence, can help to test the key question, *cui bono, who benefits*? Examining the data in this way is a common exercise, useful not least because it can help to forecast malign intentions. However, it can also be misleading since actors can craft IIO to appear as though an entirely different actor is the originator.

In a cyberattack, forensic evidence can usually demonstrate whether an intrusion did or didn't happen. In IIO, the picture is quite different. An adversary can plan IIO over a long period of time, seeding content on different websites, chat groups and the dark web[23], in preparation for an intensification of activities. Alternatively, the activities can be spontaneous and uncoordinated. Yet in both cases, the content and behaviour will often overlap with authentic issues in organic ways. For example, in 2018 Facebook removed a group created by an adversary to inspire a public demonstration, to which hundreds of real, genuinely engaged people signed up to participate.[24] It could hardly be claimed that the adversary was the "cause" of the issue, or that their "effect" was a demonstration. Rather, the impact of the IIO may be assessed in terms of how it skewed narratives, inserted disinformation into the debate, and contributed to an overall fracturing of trust in public debate. Exactly what is being attributed can raise complex questions, since authentic grievances are often the target of IIO.

**Open-source contextual evidence** focuses on the content of digital media: text, images, video, hashtags, narratives, and languages.

The aim is to understand how the content fits within a geopolitical context. This helps to develop an appreciation for the motivations behind the IIO, and its intended effects. Does the IIO intend to polarise debate? Undermine the credibility of an actor? Affect the will to vote or defend a territory? Contextual analysis is crucial to connecting the actor's behaviour and content to their and their audience's political context.

The report sample was analysed to identify the research methods used to investigate IIO. Contextual evidence, and in particular, geopolitical context and narrative analysis, were the two most frequently identified techniques, with one or the other found in 95 percent of reports.

- **Geopolitical context analysis**, explaining the background to politically motivated content, was identified in 83% of the reports
- **Narrative and discourse analysis** was identified in 83% of the reports
- **Hashtag analysis** was identified in 31% of reports
- **Linguistic analysis** was identified in 24% of reports[25]

An innovative source of open-data context-based attributions in the intergovernmental sector has been the East Stratcom Taskforce (ESTF), which sits in the Strategic Communication Division of the European External Action Service. ESTF maintains an archive of over 11,000 examples of disinformation that support pro-Kremlin messages. Some sources are state media, others may be considered pro-Kremlin media. Though created by the European Council to "challenge Russia's ongoing disinformation campaigns" in Europe[26] and funded by the European Parliament, the website hosting the taskforce's work has the disclaimer that it does not represent an official EU position.[27] Attribution terms used include, "the Kremlin's hostile propaganda,"[28] "Kremlin-backed disinformation" and "pro-Kremlin disinformation outlets."[29] The purpose is seemingly to signal knowledge of Russian activities and potentially deter the continuation of these activities, but not to make a high-level political statement of blame each time an example is identified. Working with open sources such as media means that attribution of some types of IIO does not depend on secret data and can be credited simply by assessing behaviour (such as communication techniques used), and contextual information (the narratives in a political context). Some governments have followed suit in creating open-source analysis units within foreign ministries or equivalent, though they typically share their assessments with a small group of peers at low levels of classification, and publish infrequently.

> Open-source contextual evidence relies primarily on visible platform data available through a public or semi-restricted API such as CrowdTangle for Facebook, the Twitter API, or a proprietary tool designed to analyse data acquired or scraped from digital platforms

Open-source contextual evidence relies primarily on visible platform data available through a public or semi-restricted API such as CrowdTangle for Facebook, the Twitter API, or a proprietary tool designed to analyse data acquired or scraped from digital platforms. This data is analysed and published as individual examples of what the IIO appears to be doing in general. Together with other information sources from the technical and behavioural categories, this can build toward a case that an IIO is being conducted by a specific actor for a specific purpose.

**Proprietary and classified data sources** have the added advantage of access to meta-data as well as to non-public or hidden accounts, pages, and groups; data that would not otherwise be accessible in open-source data collection. In rare cases, investigative journalists will infiltrate closed social media groups or chats, exchange information with companies and intelligence agencies, or exfiltrate proprietary or classified information to access additional data sources for attributions that they will make public. However, proprietary and classified sources have an indisputable advantage when it comes to assessing the widest possible dataset, including in contextual analyses.

## Legal and ethical assessment

Legal assessments refer to considerations that take place within an organisation prior to the public communication of an attribution. This informs most importantly the wording of the attribution, so as to minimise the possible liability of the attributor. However, a variety of other factors play a role in how the attribution is expressed, depending on the sources of evidence used. Open-source data is heavily dependent on the ethics of what is being published, such as whether use of data (including leaked information) is proportionate, and if the research methodologies are sound. The main legal risks are whether the attributor opens themselves up to libel laws which may differ in different jurisdictions. Researchers are at risk of being banned from platforms if their research methods contravene standards sets by the platforms, decisions that

platforms can make according to their own justifications despite the obvious conflict of interests.[30] Increasingly, individual journalists and researchers conducting investigations find themselves at personal risk.

There have been occasional open-source investigations about proprietary data sources. However, such studies are limited by research ethics, and particularly a sense of proportionality. In a series of studies by the NATO Strategic Communications Center of Excellence, researchers wanted to test social media companies' response to tip-offs about fake accounts. The authors purchased over 50,000 fake engagements at a cost of €300, and then notified the platforms of the inauthentic behaviour. The experiment showed that 80% of the accounts were still active one month later.[31] Giving funding to these social media manipulation companies is clearly ethically questionable but was weighed up against the need to better understand how the industry works in practice, given that platform figures are not independently verified. In another study, military personnel engaged in a NATO exercise were fed manipulated social media content in order to generate geolocation data.[32] By mimicking psychological operations against one's own troops, the report demonstrated the kind of technical data that digital platforms manipulated by IIO can provide an adversary.

Privacy is a dominant consideration for organisations working with proprietary evidence. Digital platforms are concerned about releasing identifiable personal data that could impact upon individual platform users who encounter IIO. GDPR is often invoked as a reason for digital platforms not sharing proprietary data; however, GDPR offers significant exemptions for research. Legal considerations are often less altruistic and can include geopolitical and commercial assessments aimed at determining the impact of an attribution on a company's interests, such as regulatory factors, advertising revenue, and access to markets.[33] It has been claimed for example that Facebook's threat intelligence team prioritises IIO in "the US/Western Europe and foreign adversaries such as Russia/Iran/etc," something that the dataset used in this study appears to confirm.[34]

Privacy is also a major concern for classified data users. Intelligence agencies usually investigate an issue because of concerns about an actor's behaviour and intentions. Attributions are shared among allied intelligence agencies so that those agencies build a similar assessment of that actor's capabilities and direct their future activities accordingly. This helps countries to develop actor-specific strategies, which take a holistic view on how to handle problematic behaviour in the international community. Sometimes, public attributions of IIO are mentioned in statements of an actor-specific strategy or threat intelligence statements, to inform the public, as well as signal and

deter. Such attributions are also often made by governments to support a legal response, such as sanctions or expulsions.

Ambiguity is a fundamental component of attribution. For example, reporting on one of the most heavily attributed IIO sources, the Internet Research Agency (IRA), has relied on a number of obfuscating formulations.[35] The IRA is a troll farm located in St. Petersburg, Russia, allegedly owned by Yevgeny Prigozhin, who reportedly has a close relationship to Russian President Putin.[36] The IRA is not owned by the Russian government, although it is likely that it has acted on its behalf, or under instructions.[37] While this distinction may feel like semantics, it is significant as it allows Putin to distance himself from these activities. Attribution language is therefore attempting to communicate a connection between malicious actors and nation-state sponsors, without overstating that which has not yet been proven in publicly available sources.[38] Furthermore, ambiguity also protects the methods and techniques used to identify the actors.

The political fallout of an attribution can affect future attribution research. In April 2020, a European External Action Service report on Covid-19 disinformation was, according to reports, watered down following diplomatic pressure from Beijing.[39] The report summarised existing research that claimed China was running IIO and did not contain new assessments. Still, the attribution of IIO to China in an EU publication allegedly led to threats of diplomatic reprisals from Chinese representatives.

# 3. Future prospects

Each of the three information sources have strengths and limitations. These limitations become most pronounced when the sources are used in isolation. This section draws together some key questions and points to some future directions for the field.

## Limitations of open-source data holders

Open-source research is the principle means by which detailed information about IIO reaches the public. Yet, the information that reaches the public is the tip of an iceberg of undisclosed size and shape, figuratively speaking. The resulting dynamic is one characterised by in-group and out-group research. In-group researchers have the inside track on platform and government attributed IIO. These researchers tend to receive parts of their funding from industry or government, or have developed relationships of trust over time. This grants them greater exposure to the discreet worlds of proprietary and classified data. In practice, this rarely means direct access to that data. Rather, it might involve briefings, background, and nudges toward open-source examples that exemplify key activities. In some cases, privileged access to APIs or data frontends is offered. It is about access to *information* rather than *data* per se. The price of entry is a relationship of dependency: the in-group is unlikely to criticise or contradict its sponsor's assessment or challenge the structural shortcomings that favour it. In short, building trust with government and industry means accepting a status quo that ultimately damages the research community as a whole.

Although this may not sound like a beneficial arrangement for the inner circle of researchers, out-group researchers are significantly disadvantaged. Their access even to basic information is severely compromised by a lack of engagement with the data owners. This makes it harder to interpret open data sources, to understand the signals and nudges within public statements of attribution, and to compete with the frequency and timing of the analyses of the in-group. Digital platforms such as Facebook and YouTube get to decide who can study their data. The present structure governing research into IIO attribution is therefore anti-competitive and anti-knowledge. Out-group researchers who devise their own means of accessing sensitive data risk finding themselves contravening platform policies, leaving them blacklisted and marginalised within the community. This reduces the potential of research to shed light on important socio-political phenomena such as IIO, and hence

does not serve the interests of the public.

Having said this, it is important to recognise that confirming an attribution does not have to be central to open-source research. Researchers can analyse a great deal of IIO tactics, techniques, and procedures without necessarily confirming who coordinated the IIO. High-quality research into TTPs is incredibly valuable and can help to demonstrate the value of an active, vibrant, and well-informed research community to the holders of sensitive data. Researchers often complain that a lack of data hampers their ability to analyse IIO. While this may be the case, we acknowledge that the data used by a digital platform, threat intelligence company, or intelligence agency to build confidence in an attribution is unlikely to be shared outside of a small group of confidants under any circumstances. Rather than focusing on attribution data, cross sector collaboration initiatives should focus on:

- **Transparent methodologies**. Providing more transparent information about the methodology, criteria and confidence that support public attributions made by governments and companies

- **TTP-focused information releases**. Improving access to the behavioural and contextual data necessary for high-quality research into TTPs

- **Consistent formats**. The data that is provided should use consistent formats to support cross-platform analysis that adds value to whatever an individual data owner can assess

- **A focus on relative strengths**. Governments and platforms should consider how to allow access to information and data to researchers particularly in cases where open-source researchers have relative advantages over their counterparts. Learning to understand and appreciate these differences and how they can contribute to achieving mutual interests is key.

## Limitations of proprietary data holders

Interference in the 2016 US Presidential Election was a wake-up call for the digital platforms that led to a significant investment in threat intelligence capabilities. It is worth noting that the first 3 publicly announced takedowns came from Facebook, You Tube and Twitter in late 2017, with 24 following in 2018.[40] This is, in other words, a field that is still in its infancy. In this early phase, takedowns were released as public relations announcements, often with a triumphant tone. This is wholly inappropriate. Finding themselves at the centre of a web of state espionage and active measures, it could be argued that the shock of 2016 set the platforms on a course to treat their proprietary data as equivalent to classified data. There are many reasons for this, some of which are stronger than others:

- **Cultural affinity**. Many digital platforms recruit analysts from intelligence agencies and create threat intelligence units based on government models. Their working cultures to some extent mimic work done with classified sources.

- **Protecting tradecraft**. Digital platforms use a variety of techniques to identify, deter, and track inauthentic behaviour. The owners of proprietary data argue that their data should be treated as equivalent to classified sources since publication risks revealing how it was collected and hence provides the adversary with the means to evade future detection.

- **Data privacy concerns**. Digital platforms have a responsibility to protect user data. Many authentic users may be caught up in an IIO, and their personally identifying data should be treated with discretion. It is not always straightforward to access, isolate, and share data in a useful format and timely fashion. Importantly, it is not always possible to guarantee protection of identifying personal data when data is shared. Furthermore, compliance with regulation such as GDPR and the patchwork of US data privacy laws provides a layer of complication that leads to a conservative posture (albeit one that seemingly exaggerates the constraints of regulation when it is in platform interests to do so).

- **Commercial sensitivity**. Many digital platforms rely on selling bulk datasets about their users to the advertising industry. Furthermore, their algorithms and other product features are commercially sensitive. It is therefore against a company's commercial interests to share data and reveal information about how their backends function because it may offer advantages to competitors. IIO data that infringes on commercial sensitivities should therefore be treated as equivalent to classified information, according to the owners of the data.

- **Protecting the business model**. Public understanding of how digital platforms use their data for commercial purposes is low, particularly in relation to free-to-use services. Scandals such as Cambridge Analytica demonstrate that increased transparency about business models is likely to drive users off platforms. Furthermore, it is in digital platforms' interests to present IIO as individual anomalies rather than systemic flaws in the way their platforms function. In other words, withholding information about IIO, and only releasing small portions of it piecemeal or as aggregated data, is a way of protecting the business model.

It is fundamentally in the interests of the owners of proprietary data to only release as much technical and behavioural data as they are compelled to do by law. When

In many countries, creating an open-source analysis capability outside of the intelligence agencies has been an important step. Since open-source data collection still involves governments collecting data on groups and individuals, processes remain governed by legal frameworks for intelligence collection.

it comes to the most sensitive data used in attributions, it is simply not realistic to envisage significant amounts of data being released into the public domain. Rather, the focus should be on improving the ability of data holders to increase the consistency and transparency of their methodology and confidence assessments. They should be able to share better information regardless of data issues. If data holders used a common framework in their attributions and made at least some high-level statements about how sources contribute to a confidence level, information about attributions could be compared across platforms and cases to better understand IIO. If the open-source researchers who confirm platform assessments could be assured of independence, with no favouritism or limits to who can review such information, the processes would better serve the public interest.

## Limitations of classified data holders

Governments have also been slow to respond to the threat of IIO. Legal frameworks are barely adapted to the digital era, much less to evolving trends at the cutting edge of the social media influence industry. Efforts to build counter-IIO capabilities vary country to country; but still, there are relatively few examples where governments publicly attribute IIO, even as they are becoming more confident about attributing cyber-attacks. In many countries, creating an open-source analysis capability outside of the intelligence agencies has been an important step. Since open-source data collection still involves governments collecting data on groups and individuals, processes remain governed by legal frameworks for intelligence collection. The advantage is that data is more readily used in reports at lower classification levels, which in practice means it can be circulated among allies, and occasionally published or leaked. More significant structural problems still hamper government attributions, however:

- **Political will**. Politicians are rarely trained in counter-IIO and its nuances. Decision-making is challenging even when confidence is high. The decision to attribute is ultimately political, and

hence can be affected by day-to-day political matters. Events facing the politician completely unrelated to the issue might impact a decision. Equally challenging is the potential response from a named actor, especially if it is a country. It may be expected that an accusation of guilt has consequences, often involving whatever diplomatic levers are at hand. In other words, the political will to attribute can place at risk ongoing shared interests with the attributed country, despite those issues having no connection to IIO.

- **Priorities**. Intelligence agencies need a legal reason to investigate IIO. More importantly, with limited resources the IIO also needs to represent a high priority threat to national security. Given the range of threats intelligence agencies must prioritise, in practice this means focusing on a limited number of persistent adversaries during specific periods such as elections.

- **Protecting tradecraft**. Protecting tradecraft and personal data weigh heavily on the ability to share information. However, while the exposure of tradecraft would provide a technical inconvenience for digital platforms, for intelligence agencies the ramifications can cost lives. It is therefore unlikely that IIO would provide sufficient motivation to take risks, and hence it is unlikely that a government attribution of IIO would reveal any technical data whatsoever.

- **Overcoming stovepipes**. Finally, intelligence agencies have a long tradition of relying on stovepipes to protect information. IIO, however, demands multiple skillsets. This means more than each unit simply working to its strengths and then compiling an assessment; it requires collaboration at all stages. IIO has not traditionally been a focal point of intelligence analysis and the organisational structures are often not finetuned to produce joined-up work for this challenge.

## Enhancing cooperation

In our view, there are two opportunities for improving the state of IIO attribution. The first is in a community framework that makes transparent the building blocks of an attribution. We have outlined, and to a certain extent demonstrated, the relevance of such a framework in this report. The matrix spanning three areas of the information environment (open, proprietary, and classified) and four types of evidence (technical, behavioural, contextual, and legal/ethical) enables a logical breakdown and classification of types. From this starting point, more granular structured analysis of TTPs can take place. It is clear that there is a relationship of symbiosis between the layers; yet, many of the current practical working structures remain

opaque and overall do not serve the public well.

For the second, it seems that the clearest areas of collaboration are around behavioural, contextual, and legal-ethical assessments. Technical evidence will always be peculiarly sensitive. In our view, it is not realistic to expect *data* sharing in this area. Rather, the field needs better *information* sharing. If the holders of classified and proprietary technical and behavioural evidence want to present their attributions as the most apolitical and objective assessments possible, they need open-source researchers to get better at what they do.[41] This is only possible with a firm commitment to information-sharing, transparency, and honesty. Developing a consistent terminology to describe TTPs, and best practice to support quality in the field, is the next crucial step in making IIO attribution more precise and comprehensible to a public still largely in the dark about the threat IIO offers.

> A search on the company website reveals a relevant connection to a contract that the PR company holds. With the suspicion that many of the accounts are using misleading identities in support of an IIO, the analyst builds a dossier of information

# Annex A: Three archetypal IIO attributions

In order to exemplify how these processes typically play out, we have outlined three archetypes: one led by open-source methods, one by proprietary sources, and one by classified sources. These examples demonstrate that there is a clear division of labour where different kinds of expertise each have a role to play. It is also clear that organisations like to set up protective structures. For example, governments seem to prefer to allow where possible digital platforms and open-source researchers to publish the information. Likewise, platforms prefer to use open-sources to establish public-facing evidence. In both cases, this is no doubt because countries – and even digital platforms – want to appear to the public as neutral, reliable arbiters rather than political players.[42]

[H], [M], and [L] below refer to confidence levels: high, medium and low respectively.

Scenario 1: In an **open source-led attribution**, a journalist or researcher comes across a closed social media group that mixes disinformation about vaccines with other legitimate health information. They join the group under a false identity and take screenshots of the group's content and user account details. They analyse the content and make assessments of the intentions of the group. They run searches on the usernames and attempt to find evidence of who is behind the accounts. They look for cross-platform activities, and follow-up on links to websites. Several accounts seem to be tied to a web domain owned by a PR company in a given country. A search on the company website reveals a relevant connection to a contract that the PR company holds. With the suspicion that many of the accounts are using misleading identities in support of an IIO, the analyst builds a dossier of information and reports on it both in a publication and to the digital

platform owner. The overall assessment could be expressed in a confidence interval that acknowledges gaps in the data and methodological weakness.

| | Technical evidence | Behavioural evidence | Contextual evidence | Legal & ethical assessment |
|---|---|---|---|---|
| **Open source** | Web domain ownership, IP addresses, economic ties [H] | Account activity, page activity, posting/ cross-posting, sharing, follows, network map [M] | Media content, discourse and narratives, linguistics, political context [M] | Public interest attribution |
| **Proprietary source** | n/a | Some non-public data accessed [H] | Some non-public data accessed [M] | May contravene platform terms of service |
| **Classified source** | n/a | n/a | n/a | n/a |

Scenario 2: In a **proprietary source attribution**, a digital platform is tipped off about some suspicious looking accounts that are posting links to alternative news sources that appear to be clickbait. Technical analysis reveals dozens of accounts set up by the same user in a systematic manner. Behavioural analysis also suggests coordination with at least four other networks based in the same country and doing similar things. Prior to removal of the accounts and content, the platform advises a research team of the issue. The research team finds public examples of the activities and writes a short report detailing some behavioural and contextual evidence. The platform announces a takedown of accounts and pages spreading commercially motivated disinformation, and the research report is published immediately after. The attribution is high confidence, based on proprietary technical and behavioural evidence.

| | Technical evidence | Behavioural evidence | Contextual evidence | Legal & ethical assessment |
|---|---|---|---|---|
| **Open source** | n/a (inherited from platform) | Examples of account activity, page activity, posting/cross-posting [H] | Examples of media content, discourse and narratives, other contextual information [M] | Platform transparency |
| **Proprietary source** | IP addresses, geolocation [H] | Account activity, page activity, posting/ cross-posting [H] | Low relevance | Terms of service |
| **Classified source** | n/a | n/a | n/a | n/a |

Scenario 3: In a **classified source attribution**, a request is made to an intelligence agency to analyse the behaviour of an actor suspected of electoral interference. The agency considers its technical sources and is able to deliver relevant data on the online activities of a number of individuals working for the suspected organisation. The assessment is high confidence. This classified information can inform the government's strategy toward the actor in question as well as efforts to protect the election. Intelligence is shared with allies. Meanwhile, open-source analysts within government attempt to see to what extent the case can be shared with the public without compromising technical sources. This information could support a high-level public statement of risk about general threats to the election. The threat could also be shared with independent researchers and digital platforms, which would set in motion further open-source research.

| | Technical evidence | Behavioural evidence | Contextual evidence | Legal & ethical assessment |
|---|---|---|---|---|
| **Open source** | How much evidence can be found with purely open sources? | How much evidence can be found with purely open sources? | Data shared as examples of likely interference | Assessment of risk of electoral interference announced to public |
| **Proprietary source** | Can technical evidence be acquired by warrant? | n/a | n/a | Platform can be informed and asked to intervene |
| **Classified source** | Technical evidence tying actor to activity [H] | Full insight into user activity [H] | | Informs actor-specific strategy; evidence cannot be revealed without losing access to organisation computers |

# Annex B: Findings from the literature review

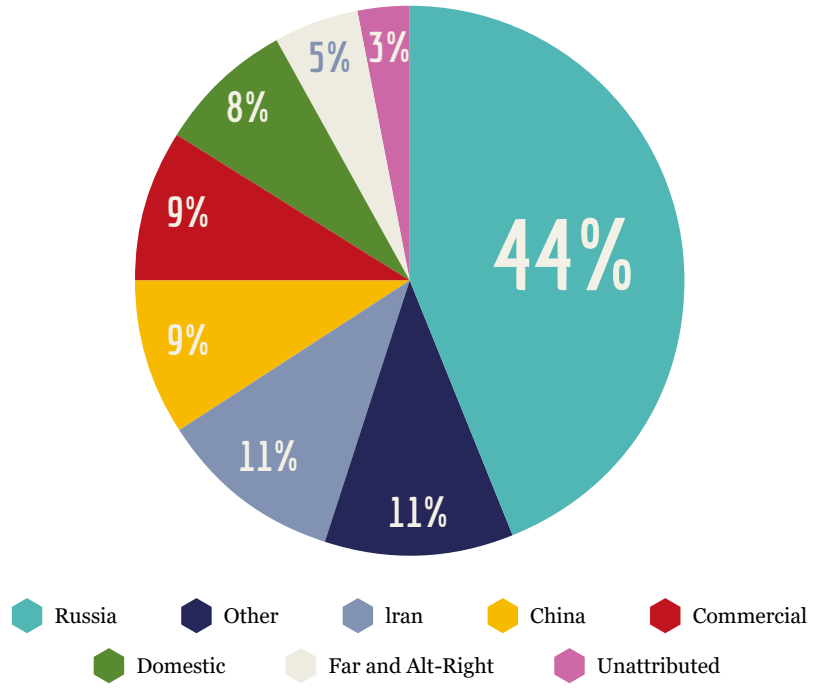


Figure 1: Distribution of Attributions[43]

Research conducted by DFRLab's Foreign Interference Attribution Tracker (FIAT), which records allegations of interference in the US 2020 Presidential elections, shows a steep increase in the number of public attributions made since 2018.[44] The FIAT methodology also highlights that not all the allegations recorded in their dataset score highly when assessed on their credibility, objectivity, transparency or evidence.

Figure 1 above shows the overall breakdown of the actors publicly named in IIO attributions. It may be tempting to believe that this represents the overall threat environment; however it is more likely that familiarity with the TTPs of actors such as Russia, alongside legal precedents in naming them, and strategic prioritisation of focusing on these threat actors, has heavily skewed this distribution.

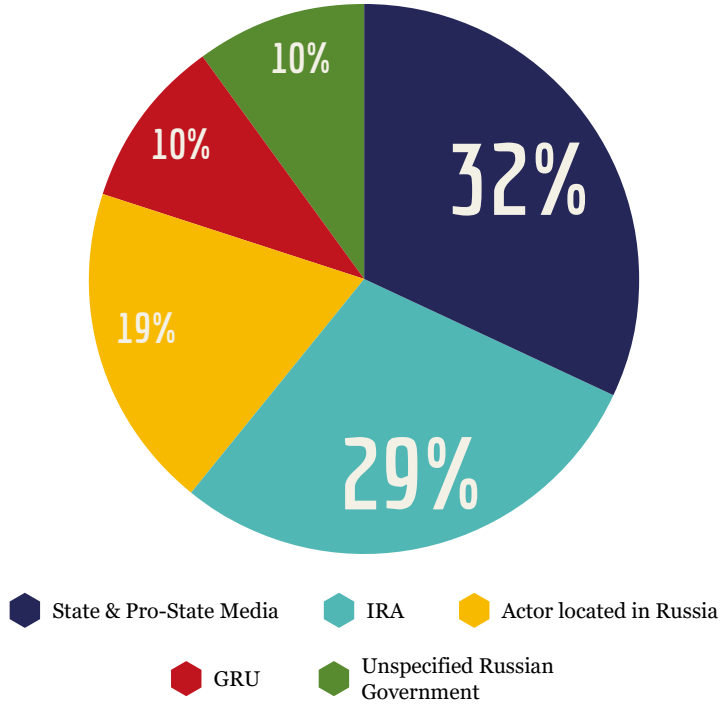


Figure 2: Breakdown of Russia-based Actors[45]

Figure 2 above shows the proportion of attributions made to different Russian-based actors.

The behavioural and contextual indicators of Russian activity have become familiar to IIO researchers in recent years. For example, Graphika's Secondary Infektion report[46] found predictable tactics across 2,500 pieces of content over 6 years, which they attributed to "a large-scale, persistent threat actor from Russia that worked in parallel to the Internet Research Agency (IRA) and the GRU but was systematically different in its approach." The contextual and legal indicators build on the illegal annexation of Crimea and subsequent interference in the 2016 US election among others, providing a clear political basis for attributing Russia without controversy.

Actors located in Iran were the next most frequently named actors (12%).

Iranian State Media or Pro-Iranian Media, such as the International Union of Virtual Media (IUVM) designated by the US for being owned or controlled by the Iranian Revolutionary Guards Corps -Quds Force (IRGC-QF),[47] featured prominently in these attribution assessments. IUVM is an Iranian cyber group that was first linked to

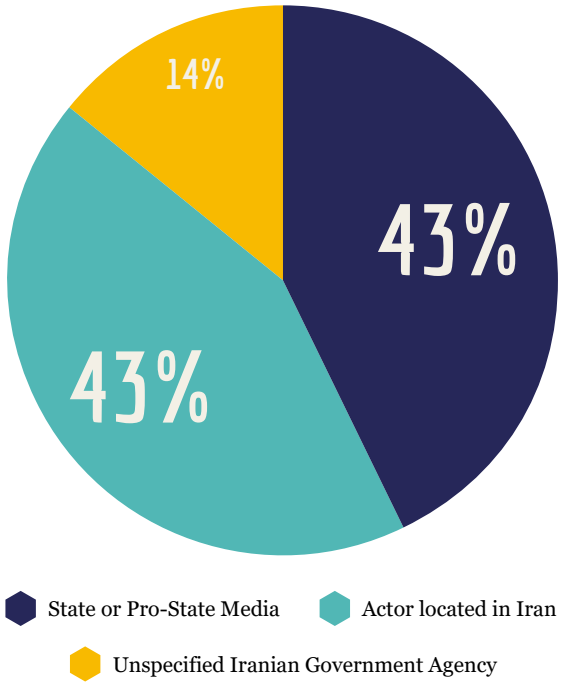


Figure 3: Breakdown of Attributions to Iran-based Actors

the Iranian government by cyber security company FireEye in 2018.[48] FireEye also used site registration data and links between social media accounts to Iranian phone numbers in their attribution in another report in this sample.[49]

The majority of attributions of actors located in China named the Chinese State, or state-backed organisations. The 8% of operations coded as domestic covered activities in Myanmar, Pakistan, Serbia, Georgia and Honduras. Other named actors included Saudi Arabia, Egypt, the UAE and Syria. The reports that did not make an attribution discussed the tactics deployed rather than who may have been responsible.

The more recent proliferation of IIO actors, including domestic actors, will provide increasing challenges to attribution tradecraft both in identifying behavioural traits, and in coping with political and ethical sensitivities. It also raises questions about whether the drive to publish an attribution, however vague, is justified or necessary and in what circumstances it may be better not to attribute.

## Most public IIO attributions are dependent on platform data

Proprietary data, which refers to the backend of digital platforms, and open data, which

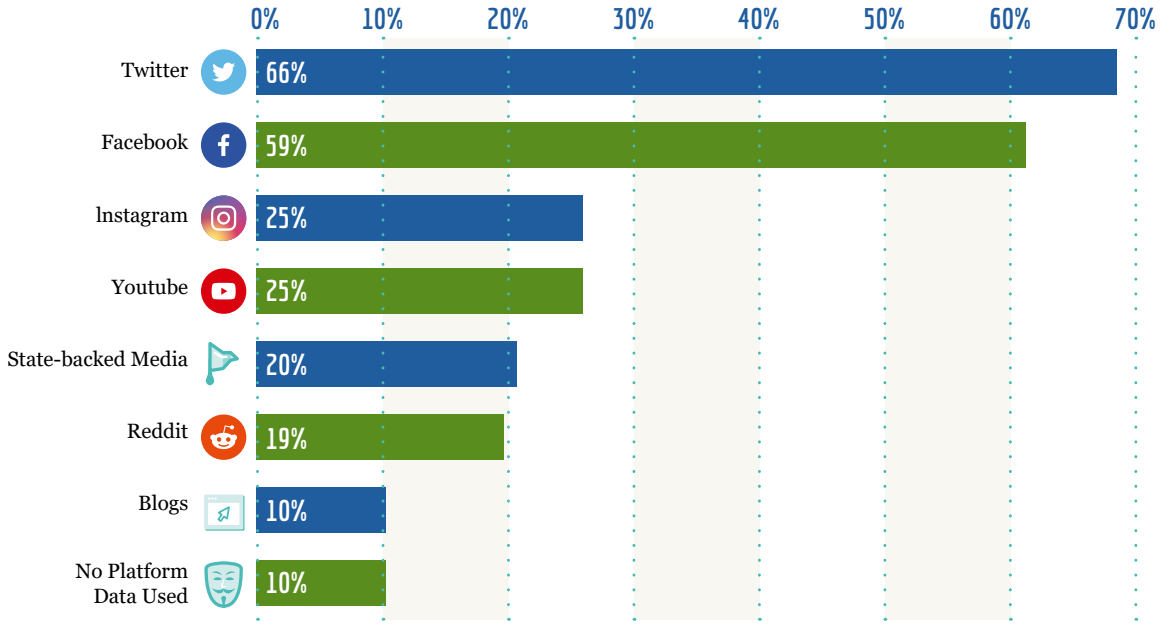


Figure 4: Data Sources

refers to data that can be accessed freely through a platform API, were significant sources of information in the reports studied. The main data sources identified are shown in Figure 4 above.

In terms of the confidence expressed in attributions, none of the reports stated or implied a low confidence, and very few explicitly stated a confidence level at all; eight expressed high confidence and 2 expressed medium confidence. The remaining reports either did not make an attribution or did not explicitly state a level of confidence. Where confidence was not explicitly stated, the attribution was presented in a factual way, including in cases where it was recognised that an attribution could not be independently verified.

This is significant because it suggests that researchers have no choice but to trust the attributions made by platforms, despite being unable to assess the methodology that led to the attribution, or the level of confidence the platform has in their assessment. It also means that the ambiguities and nuances in the initial platform takedown, such as the difference between 'actors located in country X' or 'state-backed actors', is relied on by external researchers, journalists and policymakers who often lack access to technical data to support or challenge a platform's assessment. A clearer attribution terminology and more transparency about methodology from the platforms would significantly improve the assessments shared with the public in

research and other forms of reporting on IIO takedowns.

## Most platform takedowns attribute proxies rather than governments

Platform takedowns offer a broader sense of the overall state of IIO attribution than research reports alone. They include examples where data may not be accessible to the research community, or where researchers decide not to investigate further. According to the Disinfodex database, attributions to companies specialising in public relations, marketing and strategic communication account for some 21% of attributed IIO, making it the largest single category. This is closely followed by media organisations, including both state and alternative media platforms, which account for 18% of attributions. Sources associated with Yevgeny Prigozhin, including the St. Petersburg Internet Research Agency and Wagner Group, account for 15% of all attributions in this dataset. Political parties and intelligence agencies/militaries account for 13% of attributions each, whereas governments were attributed in 7% of the takedowns. NGO/activist groups, cyber/IT firms, and the remaining groups of negligible examples make up the final attributed actors, with 4% each. It is therefore clear that the subject matter selected for research reports on takedowns is not representative of overall platform takedown behaviour.

While the attributions sometimes identify the source with some precision, there is usually no supporting technical data from proprietary sources, and few specific details which can be cross-checked. Furthermore, many of the attributions are to agents or proxies who act on behalf of an undisclosed principal. For example, the public relations companies and media houses that have been attributed are likely to have acted on behalf of a client government, political party, or private entity; it is beyond the scope of proprietary technical analysis to make such connections, which could only be confirmed by data external to the platforms, such as financial trails or private communications. Independent researchers often struggle to make sense of the methodology used to make these attributions or the degree of confidence the digital platform has in them, which in turn means that the unaccountable legal, geopolitical and commercial considerations of private sector organisations skew the entire research process around attributions.

# Annex C: Research report sample

| Institution | #Reports |
|---|---|
| Graphika | 12 |
| Atlantic Council (inc DFRLab) | 10 |
| Stanford Internet Observatory | 9 |
| NATO Stratcom COE | 4 |
| Shorenstein Center Harvard Kennedy School | 2 |
| ASPI | 2 |
| FireEye Intelligence | 1 |
| Bellingcat | 1 |
| Graphika & DFRLab | 1 |
| The Political Quarterly | 1 |
| EU Disinfo Lab | 1 |
| ISD and LSE Institute of Global Affairs | 1 |
| LSE | 1 |

| Institution | #Reports |
|---|---|
| Brookings | 1 |
| School of Media & Public Affairs, GWU | 1 |
| CSIS | 1 |
| US Senate | 2 |
| The International Journal of Press/ Politics, 2020 | 1 |
| CIMA | 1 |
| Data & Society | 1 |
| Computational Propaganda Project, University of Oxford | 1 |
| Social Media and Political Participation Lab, New York University | 1 |
| Global Engagement Center | 1 |
| Innovation in Warfare and Strategy | 1 |
| Philip Merrill College of Journalism | 1 |

# Endnotes

1 disinfodex.org; the big three refers to Google/YouTube, Facebook, and Twitter.

2 James Pamment, Howard Nothhaft, Henrik Agardh-Twetman, Alicia Fjällhed (2018) *Countering Information Influence Activities: The State of the Art*. Stockholm: MSB

3 *RESIST Counter-Disinformation Toolkit*. London: Government Communication Service

4 Annex A provides archetypal examples of typical pathways to IIO attribution.

5 See e.g. https://www.hoover.org/sites/default/files/research/docs/lin_webready.pdf or https://cyberdefensereview.army.mil/Portals/6/Documents/CDR%20Journal%20Articles/Determinants%20of%20the%20Cyber_Kostyuk_Powell_Skach.pdf?ver=2018-07-31-093725-923 or https://www.mitpressjournals.org/doi/pdf/10.1162/ISEC_a_00266 or https://www.nap.edu/read/12997/chapter/4 or https://www.rand.org/pubs/external_publications/EP68257.html

6 Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks, Journal of Strategic Studies", 2015, Vol. 38, Nos. 1–2, 4–37, http://dx.doi.org/10.1080/01402390.2014.977382

7 Office of the Director of National Intelligence, Building Blocks of Cyber Intelligence, Accessed June 14, 2020, https://www.dni.gov/index.php/cyber-threat-framework

8 Sergio Caltagirone, Andrew Pendergast and Christopher Betz, "The Diamond Model of Intrusion Analysis, 2013, Accessed June 14, 2020, https://pdfs.semanticscholar.org/dca1/9253781fbc429d85ec09e8f0f7f2ddbe7fdf.pdf?_ga=2.254620238.791791996.1592145895-821038176.1592145895

9 There are of course potential instances of where the same campaign involves both IIO and cybersecurity dimensions, and in these cases more holistic methodologies are warranted.

10 The "backend" of a computer system or application is the part that users do not access, where data is stored. The "frontend" is the part that users use to search and filter data held in the backend. The API (Application Programming Interface) enables two systems or programmes to connect to one another and is how developers and researchers gain greater access to a system backend than is possible through the frontend.

11 We reviewed 88 reports on IIO by think tanks, universities, governments, companies and other research organisations from 2017-2020. Reports that focused on general trends in IIO, or which did not focus on detailing specific IOs, were removed from the sample. This left 59 reports that were used for further analysis. The content of each report was analysed and coded to record information including: basic metadata about the report, the source of the data, research techniques used, tactics identified as used by an IIO, whether an attribution was made or inherited from e.g. a platform takedown or other research, details of the suspected actor, languages used, whether specific geographic regions or countries appeared to have been targeted and whether an assessment was made about the objectives or motives of the IIO. In total, 14 research techniques and 41 separate tactics were identified.

12 The organisations and the respective number of reports drawn from each are shown in Annex C.

13 https://disinfodex.org. 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.

14 Accessed 4 February 2021

15 Healey, J., Beyond Attribution: Seeking National Responsibility for Cyber Attacks, Atlantic Council, February 2012, https://www.atlanticcouncil.org/wp-content/uploads/2012/02/022212_ACUS_NatlResponsibilityCyber.PDF

16 "Information Warfare" and Online News Commenting: Analyzing Forces of Social Influence Through Location-Based Commenting User Typology, Social Media and Society, July-September 2017, https://journals.sagepub.com/doi/full/10.1177/2056305117718468

17 https://www.bellingcat.com/resources/2020/12/14/navalny-fsb-methodology/

18 For example, during the 2019 Indian elections, a viral WhatsApp message claimed that the BBC had reported on a survey conducted by the CIA, KGB and Mossad that predicted a landslide victory for the Indian National Congress Party. The URL of the BBC's webpage was included in the message, seemingly to add credibility. See A Fake Survey Citing The BBC Is Back, This Time Predicting A Congress Win, Boom, 11 April 2019, https://www.boomlive.in/a-fake-survey-citing-the-bbc-is-back-this-time-predicting-a-congress-win/

19 https://cyber.fsi.stanford.edu/io/news/potemkin-pages-personas-blog

20 From Russia With Blogs, Graphika, February 2020, https://graphika.com/uploads/Graphika%20Report%20-%20From%20Russia%20with%20Blogs.pdf

21 https://www.vox.com/2019/4/19/18485535/mueller-report-redactions-data-chart

22 Inauthentic amplification describes the act of sharing content within or between platforms by networks of coordinated accounts. These accounts can be wholly or partially automated, hacked and stolen, purchased in bulk or created with avatars rather than genuine users.

23 The dark web is a part of the internet that is only accessible by means of special software that enables users to appear anonymous and untraceable.

24 https://www.vice.com/en_us/article/d3e49j/how-real-activists-learned-facebook-was-deleting-their-protest-page-for-inauthentic-behavior-russia-unite-the-right

25 An assessment of whether content was written by a native speaker or if mistakes indicate a suspected different mother tongue

26 European Council. (2015). European Council conclusions on external relations (19 March 2015). Press release 134/15. Retrieved from http://www.consilium.europa.eu/en/meetings/european-council/2015/ 03/19-20/

27 https://euvsdisinfo.eu/

28 EU prepares itself to fight back against hostile propaganda, EEAS, 14 March 2019, https://eeas.europa.eu/delegations/ukraine_id/59609/EU%20prepares%20itself%20to%20fight%20back%20against%20hostile%20propaganda

29 EEAS special report update: short assessment of narratives and disinformation around the covid-19/coronavirus pandemic, EUvsDisinfo, 24 April 2020, https://euvsdisinfo.eu/eeas-special-report-update-2-22-april/

30 https://blog.mozilla.org/en/mozilla/news/why-facebooks-claims-about-the-ad-observer-are-wrong/

31 https://www.politico.eu/article/social-media-inauthentic-behavior-google-facebook-twitter-nato-stratcom/; https://www.stratcomcoe.org/black-market-social-media-manipulation

32 https://www.stratcomcoe.org/current-digital-arena-and-its-risks-serving-military-personnel

33 Sheryl Sandberg and Top Facebook Execs Silenced an Enemy of Turkey to Prevent a Hit to the Company's Business, Propublica, https://www.propublica.org/article/sheryl-sandberg-and-top-facebook-execs-silenced-an-enemy-of-turkey-to-prevent-a-hit-to-their-business

34 Julia Carrie Wong, How Facebook let fake engagement distort global politics: a whistleblower's account. The Guardian, https://www.theguardian.com/technology/2021/apr/12/facebook-fake-engagement-whistleblower-sophie-zhang

35 "Russia's Internet Research Agency" and "The Kremlin-backed Internet Research Agency"; "The Russian government, operating through the Internet Research Agency"; "Russian information operatives working for the Internet Research Agency". See How Russia's Troll Farm Is Changing Tactics Before the Fall Election, New York Times, 29 March 2020, https://www.nytimes.com/2020/03/29/technology/russia-troll-farm-election.html, New reports detail sophistication of Russian influence efforts in U.S., CBS News, 17 December 2018, https://www.cbsnews.com/news/new-reports-detail-sophistication-of-russian-influence-efforts-in-u-s/, Senate Report: Russians Used Social Media Mostly To Target Race In 2016, NPR, 8 October 2019, https://www.npr.org/2019/10/08/768319934/senate-report-russians-used-used-social-media-mostly-to-target-race-in-2016?t=1588331552021

36 Inside the Russian Troll Factory: Zombies and a Breakneck Pace, New York Times, 18 February 2018, https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html; Yevgeny Prigozhin: who is the man leading Russia's push into Africa?, The Guardian, 11 June 2019, https://www.theguardian.com/world/2019/jun/11/yevgeny-prigozhin-who-is-the-man-leading-russias-push-into-africa

37 US Senate Select Committee on Intelligence's report on Russian Active Measures, Campaigns and Interference in the 2016 US Election, Volume 2: Russia's Use of Social Media with Additional Views, Published October 2019, https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf

38 https://about.fb.com/news/2018/07/removing-bad-actors-on-facebook/

39 https://www.politico.com/news/2020/04/25/china-put-pressure-on-eu-to-soften-coronavirus-disinformation-report-207797

40 https://disinfodex.org/

41 https://www.lawfareblog.com/intelligence-communitys-role-countering-malign-foreign-influence-social-media

42 https://www.lawfareblog.com/intelligence-communitys-role-countering-malign-foreign-influence-social-media

43 The total exceeds 100% as some reports attributed multiple actors.

44 Foreign Interference Attribution Tracker, DFRLab, Accessed 10 March 2020, https://interference2020.org

45 The total exceeds 100% as some reports attributed multiple actors

46 https://secondaryinfektion.org/report/secondary-infektion-at-a-glance/

47 https://home.treasury.gov/news/press-releases/sm1158

48 https://www.fireeye.com/content/dam/fireeye-www/current-threats/pdfs/rpt-FireEye-Iranian-IIO.pdf

49 https://www.fireeye.com/blog/threat-research/2018/08/suspected-iranian-influence-operation.html


Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org