ISD | Institute for Strategic Dialogue

Disinformation Starter Kit

The 101 of Disinformation Detection

Carl Miller
Chloe Colliver

Disinformation Starter Kit

The 101 of Disinformation Detection

Contents

Not every organisation can or should become a disinformation detective. But disinformation can threaten the activities, objectives and individuals associated with civil society groups and their work. Disinformation tactics and the responses in place to try to mitigate them online are changing rapidly. Organisations witnessing or targeted by disinformation therefore require a baseline understanding of the threats posed by disinformation and how to spot them while conducting their work. This toolkit sets out simple steps to do so.

The toolkit lays out an approach that organisations can undertake to begin to track online disinformation on subjects that they care about. The process is intended to have a very low barrier to entry, with each stage achievable using either over-the-counter or free-to-use social listening tools. For a deeper explanation of the methods, teams and skills required to build a disinformation detection system, see ISD's accompanying roadmap for the disinformation research sector: 'Developing a Civil Society Response to Online Manipulation'.

About ISD's Digital Analysis Unit

ISD's Digital Analysis Unit combines social listening and natural language processing tools with leading ethnographic research to better understand how technology is used by extremist and hateful groups. We use commercial tools that aggregate social media data to analyse broad trends in discussion and how they may be influenced by hateful groups and disinformation. Using tools co-developed by ISD, we are able to analyse specific types of hateful speech online and trace where this speech comes from. We use these insights to help policymakers and companies craft informed policy responses to hate and disinformation, and to help communities mount responses at the local level.

This report was produced with support from Luminate.

Getting Started

The toolkit provides the basic steps you need to begin tracking online disinformation on the subjects you care about

Digital technology has enabled engagement with politics and information for more people in more places than ever previously possible. Social media and communication technologies allow individuals and institutions across the globe to connect and communicate. But they also allow, and in many ways encourage, them to compete: digital information has, predictably, become yet another terrain for the power games of states, as well as an array of non-state activists, extremist groups, private companies, high-net-worth individuals, and criminal elements vying for influence.

The reality of identifying and exposing disinformation is complex and goes beyond elections and beyond Russia. The definitions, methods, tools and outputs of disinformation research remain in flux and are rapidly evolving with every new crisis or election cycle. To keep pace with the efforts of the increasing number of states exploiting technology to spread false information, adopt false identities or confer false popularity or outrage, the research community needs to share lessons and tools with partners across civil society to transform the scope, accuracy and capacity of disinformation research.

Not every organisation can or should become a disinformation detective. But disinformation can threaten the activities, objectives and individuals associated with civil society groups and their work. Organisations witnessing or targeted by disinformation therefore require an ongoing understanding of the threats represented by disinformation and how to spot them. This toolkit sets out simple steps to do so, laying out a process that organisations can undertake to begin to track online disinformation on subjects that they care about. The process is intended to have a very low barrier to entry, with each stage achievable using either over-the-counter or free-to-use social listening tools.

The toolkit has two objectives: to explain how to detect and analyse examples of false information (‘disinformation’), and to explain how to detect and analyse examples of false behaviour (‘platform manipulation’). Both require careful definitions, preparation and a very critical eye. Detecting false behaviour is an entire emerging field of research in itself, and one where researchers struggle to find generally accepted definitions and methods. Therefore anyone attempting to detect both false information and false behaviour should beware of the intricacies and grey areas of online research. After all, disinformation about disinformation is still disinformation.

TOOLKIT OBJECTIVES

- To explain how to detect and analyse examples of false information (‘disinformation’)
- To explain how to detect and analyse examples of false behaviour (‘platform manipulation’)

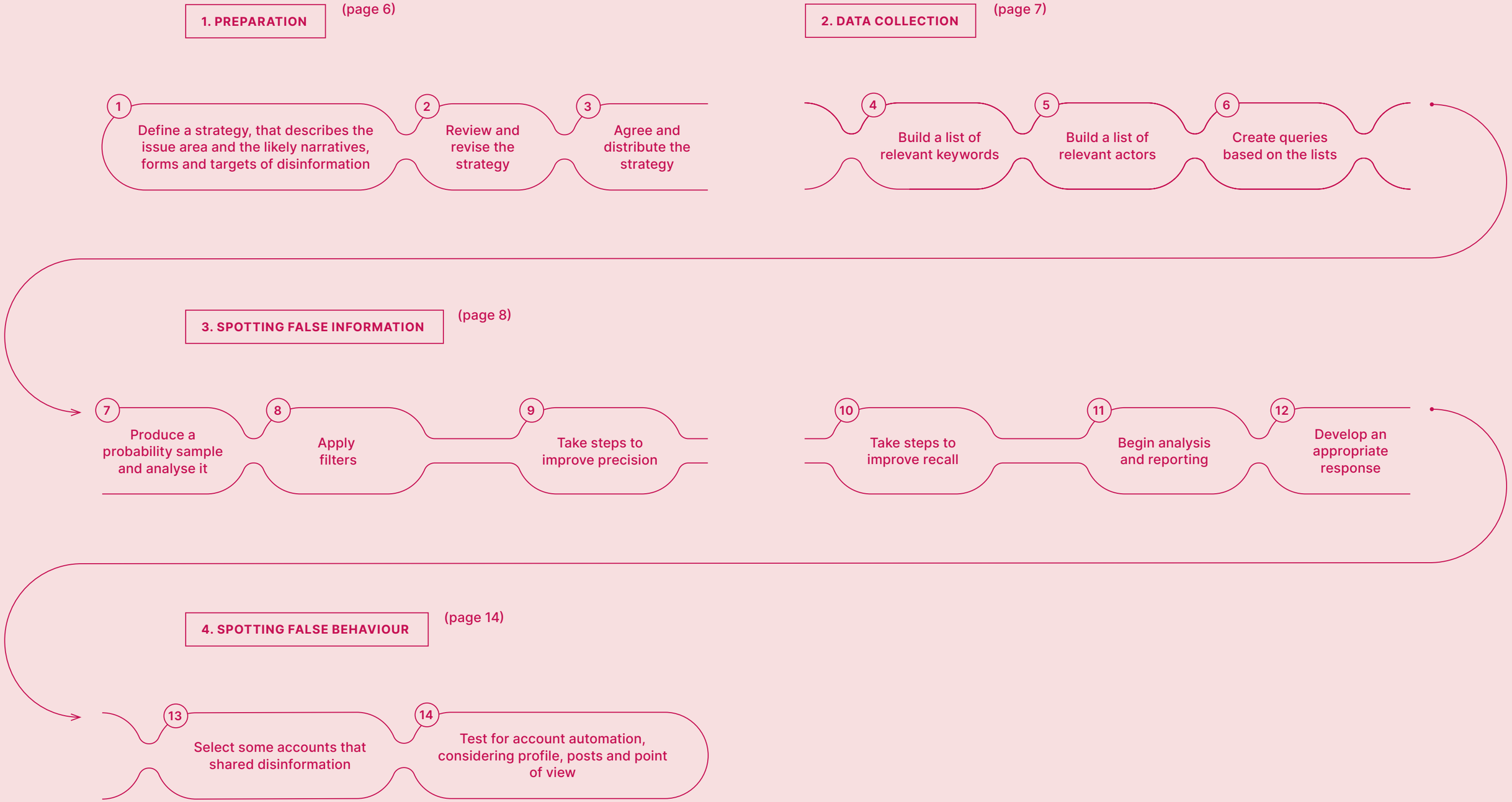
These are the steps to begin to track online disinformation.

Preparation 1

Data Collection 2

Spotting False Information 3

Spotting False Behaviour 4


1. PREPARATION
(page 6)
1
Define a strategy, that describes the issue area and the likely narratives, forms and targets of disinformation
2
Review and revise the strategy
3
Agree and distribute the strategy
2. DATA COLLECTION
(page 7)
4
Build a list of relevant keywords
5
Build a list of relevant actors
6
Create queries based on the lists
3. SPOTTING FALSE INFORMATION
(page 8)
7
Produce a probability sample and analyse it
8
Apply filters
9
Take steps to improve precision
10
Take steps to improve recall
11
Begin analysis and reporting
12
Develop an appropriate response
4. SPOTTING FALSE BEHAVIOUR
(page 14)
13
Select some accounts that shared disinformation
14
Test for account automation, considering profile, posts and point of view

1

Preparation

With online research, it is extremely easy to collect datasets that are far too large and complex to be easily handled at a click of a button. Before any technology is used, it is really important that any organisation has a very clearly defined idea of what the problem really is that they want to detect and counter.

The organisations that attempt to counter disinformation must first draw up an overall strategy document that clearly describes:

- the issue area, as tightly defined as possible
- the narratives within that issue area that are vulnerable to disinformation
- the forms of disinformation with which the organisation is concerned
- the possible or likely origins of this disinformation
- the possible or likely targets of this disinformation.

This document can be reviewed and revised, but it serves as the key orientation around which the organisation will collect data, analyse it and produce outputs. The contents should be agreed and distributed to everyone collaborating on disinformation before moving onto the next stage.

For more information about definitions for disinformation, see ‘Going Deeper on Definitions’ in Section 5 of this toolkit.

When getting prepared have you...

1. Defined a strategy, that describes the issue area and the likely narratives, forms and targets of disinformation
2. Reviewed and revised the strategy
3. Agreed and distributed the strategy

2

Data Collection

Social media data can be collected according to two broad criteria:

It contains one or a given number of keywords (which can include hashtags, numbers, links and so on).

or

It was sent by a given account, page, website, group or channel.

To collect social media data relevant to the definitions the organisation has identified, as set out in Section 1, ‘Preparation’, they need to build:

list of key words that are contained within the form(s) of disinformation with which they are concerned, related to the issue areas which are relevant, and that avoid any more general terms that will return a much larger body of data.

and

lists of actors who they know create and share disinformation related to their definitions above.

The organisation can then make ‘queries’ – or collect relevant data. Over time, the number of queries an organisation makes can increase, but to start they should create queries based on the keywords or actor lists they have generated.

These queries can be made using a variety of entry-points to data. This includes the use of social media analysis tools, often called ‘social listening tools’, which usually require paid subscriptions to allow access to stored datasets from social media or open web content. Some free tools enable limited access to similar datasets and analytics functions. Other options include data made possible through application programming interfaces (APIs). Alternatively, partnerships with research organisations or academic institutions can help provide access to large social media datasets.

The pros and cons of these various approaches to data collection are outlined in Table 1.

When collecting data have you...

4. Built a list of relevant keywords
5. Built a list of relevant actors
6. Created queries based on the lists

TABLE 1

1. Subscription online analytics tools

Pros

In-built visualisation: provide easy-to-read presentations of data

Low technical barrier to entry: interfaces built for commercial purposes are easy to use

Technical support: customer service teams help run or interrogate queries

Historic data access: some access to historic data for open sites (blogs, forums) and limited social media (Twitter)

Cons

'Black box' technology: often little transparency on how analytics are calculated and accuracy of claims; no control over changes to data access or analytics functions

Cost: even with discounts for non-profits, subscriptions can be expensive, often prohibitively so for small organisations

Limited range of social media data sources: data sources often selected on the basis of their relevance to the marketing/advertising sectors

Examples Brandwatch↗, Talkwalker↗, Meltwater↗

2. Free to access online analytics tools

Pros

In-built visualisation: often provide easy-to-read snapshots of the reach or engagement with keywords, websites or articles

Low technical barrier to entry: provide easy-to-use interfaces to query some platform APIs or keyword searches for large structured media datasets

Cost: free to use, at least through demo or limited subscription accounts

Cons

Limited range of social media data sources: limited social media data access, which is based on platform decisions and associated costs, including strict limits on historical data

Inflexible data queries: few options for querying data, offering little flexibility beyond keyword or link searches

Examples CrowdTangle link checker↗, One Million Tweet Map↗, Media Cloud↗, Information Operations Archive↗, Alliance for Securing Democracy: Hamilton 2.0 Dashboard↗, YouTube DataViewer↗

3. On-platform search functions

Pros

Low technical barrier to entry: simple interfaces for queries and results

Cost: free to use

Broad range of social media data sources: enable some visibility on any platform with a search function

Cons

Unstructured data: often no way to export data from search functions on platforms, disabling further analysis

No data visualisation: because direct data access

Limited historic data access: limited social media data access, which is based on platform decisions, including limits on historical data

Examples Twitter Advanced Search↗ and TweetDeck↗, Facebook Search↗, YouTube Search↗, Facebook ads library dashboard↗, Google search↗ (including advanced operations↗)

4. Platform APIs

Pros

Historic data access: some access to historic data for open sites (blogs, forums) and some social media (Twitter, CrowdTangle for public Facebook data)

Structured data: include structured metadata for each post or account, enabling comparative analysis on each platform

Cons

High technical barrier to entry: often require at least basic coding capability to use

No data visualisation: because direct data access, only structured raw data

Examples YouTube API↗, Twitter API↗, CrowdTangle API↗

5. Partnerships with academic or research institutions

Pros

Technical support: possibility for technical support to help run or interrogate queries

Low technical barrier to entry: can use existing infrastructure for API access and analytics

Cons

Cost: partnership establishment takes significant resource of funding and time

Examples BBC disinformation partnerships with ISD↗, Bellingcat and Australian Strategic Policy Institute↗

6. Crowdsourced data

Pros

Broad range of social media data sources: enables some visibility on messaging apps and closed platforms and removes some boundaries on data access set by platforms or vendors

Public engagement: enables visibility on potential disinformation that reaches members of the public

Cons

Noisy data: no control over the validity or accuracy of reports from volunteers

Unstructured data: requires triage and classification to make data comparable across platforms, submissions and types (image, text, video)

Limited technical support: requires additional analytics to ascertain scale or source

Limited historic data access: reliance on live real-time reporting of disinformation limits access to historic data

Examples WhatsApp reporting in Spain (Avaaz)↗, WhatsApp monitor↗ (Brazil, India, Indonesia), Comprova project, Brazil↗ (First Draft), CDR Link↗

Spotting False Information

1. Getting a Sense of Scale

The datasets collected are likely to contain some instances of disinformation, and some that do not include disinformation. There are two ways for researchers to find out how large a problem disinformation is:

1. produce a probability sample and analyse it manually
2. apply filters.

Produce a probability sample and analyse it manually

Analysts should create a random sample of the social media data they have collected by downloading the data into Excel, and using a random number generator (1 to n, where n is the number of rows in the dataset).

It is possible to use a statistical formula to ensure that the size of the probability sample is representative of the collected dataset overall, but a sample size of 100 pieces of social media content is a good rule of thumb.

An analyst can now look at the sample, and mark each post as 'disinformation' or 'not disinformation', and indeed assign additional, more specific, categories of disinformation once identified. This will show the organisation how many false positives they have collected, and by extrapolating the number of true positives found in the sample by the total data collected, analysts can estimate how many pieces of disinformation they have collected overall.

Apply Filters

Analysts can narrow the search by using some additional filters based on keywords. First, they should apply the keyword list created for the keyword data collection as a filter for the information produced by their disinformation actor list. This should remove some of the false positives from that list. Second, if they have collected a very large quantity of irrelevant information, they can create a filter to remove obviously irrelevant words.

2. Trial, Error and Refinement

There are two questions for analysts to consider when trying to monitor disinformation online relating to recall and precision:

1. 'Of the disinformation that exists, how much of it am I collecting?' (recall)
2. 'Of what I am collecting, how much is disinformation?' (precision).

Recall and precision can often exist in tension with one another: very high precision can mean that fewer examples are found; likewise an extremely high recall can introduce higher levels of noise. Analysts must make case-by-case judgements about their requirements regarding recall and precision. For instance, if they wish simply to find examples of disinformation related to their subject area, a higher precision rate will decrease the amount of time an analyst has to scroll through messages that are collected. However, if they wish to make more general estimates of the total scale of disinformation, it is also important for analysts to improve their recall to avoid systematic and consistent underestimations. Staff in organisations can improve recall and precision of their data by following a series of steps.

Steps to Improve Precision
If there are a very large number of false positives, the organisation should change their actor and keyword lists. The analyst should attempt to find whether there are certain keywords that they have used that are present across many of the false positives. If they find these words, they should be removed from the data collection queries. False positives can be found by working through the following steps:

1. Create a new probability sample of data the organisation has collected.
2. Mark a sample into the categories 'disinformation' (relevant) and 'not disinformation' (irrelevant) to find out the ratio between them.
3. Using either the social listening platform, free corpus linguistics software or online word cloud tools, measure the frequency of each word occurring in both the relevant and irrelevant categories.
4. Filter the most frequently occurring words by the keyword collection list.
5. Find words which are part of the keyword collection list, which occur frequently in content in the irrelevant category, and less frequently in the relevant category. Remove them from the keywords list for data collection.
6. Re-initiate data collection with the reduced keyword list.
7. Carry out this process iteratively over time, to refine the list of keywords being used to collect data.

Steps to Improve Recall
Recall is more difficult to measure, as it can only be established by knowing the amount of disinformation that the system has not collected. Analysts should rely on their knowledge and network of the area which disinformation may affect. For example, if they receive information that disinformation has increased or changed in nature, and this has not been picked up in the data they are collecting, they should conduct the following steps:

1. Add a number of new candidate keywords that reflect or represent false negatives, the kinds of disinformation that is not currently being collected.
2. Allow new data to be collected.
3. Create a random probability sample of the newly collected data.
4. Place each post in the sample manually into the categories 'disinformation' (relevant) 'not disinformation' (irrelevant) to find out the ratio between them.
5. Using either the social listening platform, or free third-party corpus linguistics or word cloud tools, measure the frequency of each word added in both the relevant and irrelevant categories.
6. Consider adding those keywords which occur more frequently in the 'relevant' than 'irrelevant' category as new keywords.
7. Carry out this process iteratively over time, to refine the list of keywords being used to collect data.

3. Analysis

Once analysts are comfortable with the precision and recall of their monitoring, they can report on the disinformation that they see. There are a number of outputs they might consider:

Changes over time
Using the social listening platform, analysts can create a time series showing the volume of data they have collected over time. This can show whether the scale of disinformation has increased or decreased.

Event-specific analysis
The organisation can analyse a specific time window to look for disinformation that occurs immediately after an event that is important to their issue area.

Qualitative narratives
Once they have achieved sufficiently high precision, analysts can appraise the examples of disinformation manually. They can write summaries of what they see, breaking the disinformation down into different meaningful categories based on actor-type, false claim, issue area.

Thresholds and alerts
Some social listening tools allow users to implement a series of alerts in certain circumstances, for instance, if the data collection exceeds a particular volume. These alerts can be used to set thresholds to notify the organisation automatically when surges in disinformation occur that exceed the average quantities over, for instance, a given day or week.

4. Reporting

Groups may wish to understand disinformation targeting them in order to better equip communities to be resilient to its dangers. This sometimes requires alerting people to the presence of disinformation narratives, incidents or behaviours. If communicating to others about disinformation, there is a key underlying principle that should be adhered to: avoid undue amplification.

When considering media reporting, public awareness raising or alerts to communities or organisational memberships, groups must think about the potential impacts of the amplification of false or misleading content. It is important to bear in mind the responsibilities of the researcher.

Kwan's report for Frist Draft, Responsible Reporting in an Age of Information Disorder↗, (Kwan, 2019), outlines this tricky balancing act:

> *When confronted with mis- and disinformation, your first impulse may be to debunk: bring the falsehood into the light, tell the public what's going on and explain why it is untrue. When coverage… is the end goal of many disinformation agents, however, sunlight may not always be the best disinfectant.*

Kwan raises some instructive guiding questions for researchers considering highlighting, flagging or exposing disinformation:

Why are you exposing attempts at disinformation?
- To educate the public about disinformation campaigns so that they can be more vigilant?
- To try and encourage technology companies or governments to act?

What should the appropriate response be?
- Should it focus on debunking the content, on the actors behind the content, or on the platforms that allow the content to spread?
- How can we highlight the existence of such behaviour without perpetuating the messages that they are boosting?

When working to spot false information have you...

7 Produced a probability sample and analyse it

8 Applied filters

9 Taken steps to improve precision

10 Taken steps to improve recall

11 Begun analysis and reporting

12 Developed an appropriate response

Spotting False Behaviour

The process described in Section 3 allows analysts to identify disinformation based on the content – and therefore the claims – made in the messages themselves. However, disinformation can also describe a series of deceptive practices related to the propagation and reception of the content; making it appear more visible, more popular or indeed more unpopular than it otherwise would be. Taken together, this is called platform manipulation.

It can be extremely difficult to detect platform manipulation definitively. A number of techniques exist to camouflage its existence, and the data available to civil society organisations is often inadequate to attribute it to a given actor. It is important that analysts in organisations understand these limitations before undertaking or communicating any findings they make in appropriately caveated ways. However, there are some signals that can be used to gain a tentative sense of whether different kinds of platform manipulation have been used.

Account Automation

Automation is the complete or partial use of scripting to automate a social media account's behaviour. It can be used in combination with automated account registration to control a very large number of accounts in order to simulate the activity of a crowd. Media, research and public attention has largely focused on these automated accounts – commonly referred to as 'bots' – to try to understand platform manipulation. Many bots are harmless, and often readily label themselves overtly as bots: weather reporting bots or customer service bots, to name just a couple. But covert bots, which do not provide any transparency about the fact that they are run by scripted software and are used to imitate human interaction, are deceptive and often used to conduct platform manipulation.

Using the process outlined in Section 3, 'Initial Detection', it is possible to select accounts that have either sent or shared disinformation manually. A number of services, such as Botometer↗ and Bot Sentinel↗, exist which can give scores to accounts on their likelihood of being a bot. However, both services have been restricted to Twitter, and Twitter itself has explicitly criticised↗ their accuracy.

Instead of using these services, we recommend that analysts draw more tentative conclusions based on an impressionistic appraisal of the data. To do this, they should take a probability sample of accounts that have either sent, interacted with, or shared information which the organisation believes is disinformation.

There are a number of guides published by research organisations, such as the DFR Lab↗ and First Draft↗, which provide advice on how to identify 'bots', or possibly automated accounts. In broad terms, there are three dimensions to accounts that analysts should consider: profile, posts and point of view.

Profile

Automated accounts tend to be anonymous, although in some cases 'compromised accounts' are used that do relate to a real person, but whose control has been transferred to another actor. A reverse image search can be done on the profile picture to identify whether it is a stock image, or used by other social media accounts. Here too, however, there are no definitive indicators of automation, as many other social media users choose not to make their identity known. The time the account was created can be another indicator, especially if many accounts the analyst appraises were created over the same time period. However, 'aged' social media accounts are available for purchase in bulk, so influence operations can be conducted using recently acquired accounts that appear to be many years old.

Posts

One of the most commonly used indicators for automation is the volume of activity an account shows on a social media platform. Researchers have often set thresholds for 'inorganically' high levels of posting or sharing, for instance, as an identifier. However, this has proved to be controversial and often inaccurate as a signal of automation, as many humans behave as avidly as automated accounts on social media. Activity over time can be a valuable indicator, where an analyst spots that an account is so continuously active across the day and night that there would be no time for its operator to sleep. However, care should be taken in case a number of people control the account collectively. A very high number of shares can indicate an account is engaged in artificial amplification, but a number of human beings also use social media largely to amplify certain voices and messages.

Point of view

The dedication of an account to a single issue, theme or campaign may also be an indicator that it is engaged in platform manipulation; however, it may be simply that its user has an overwhelming interest in the issue in question. A sudden change in thematic interest can suggest that a network has been 'activated' for political effects, but likewise may simply demonstrate that an individual has become interested in a new topic. Another indicator might be that the account switches language over time, especially if linked to themes that become newly important geopolitically.

Overall, there is no single key indicator that definitely proves an account is being used for platform manipulation. Human beings can behave surprisingly similarly to bots online, and bots can be made to behave in surprisingly humanlike ways. However, the existence of multiple indicators together can collectively contribute to an overall suggestion that an account behaves in 'inorganically'. This is the most definitive that an organisation should hope to be.

When working to spot false behaviour have you...

13 Selected some accounts that shared disinformation

14 Tested for account automation, considering profile, posts and point of view

Further Tools

Research tools for studying the elements of websites are outlined below.

TABLE 2

1. Domain Ownership			
Are domain registry records available?	Element / Feature	Domain URL or internet protocol (IP) address	Note The comprehensiveness of domain registry records, particularly Whois lookups, varies by database. Best practice dictates checking multiple Whois lookups for completeness.
	Approach	Whois record lookup	
	Resources	https://viewdns.info/ http://whois.domaintools.com/ https://who.is/	
What other domains are registered at the same IP address?	Element / Feature	Domain URL or internet protocol (IP) address	
	Approach	Reverse IP domain search	
	Resources	https://www.yougetsignal.com/tools/web-sites-on-web-server/ https://viewdns.info/reversewhois/	
If the domain registrant name appears to be that of a business entity, what details can be gleaned from corporate registry sources?	Element / Feature	Registrant name or organisation	Note While open source tools such as OpenCorporates can be useful to identify leads while conducting business intelligence research, it may be worth considering searches of paid third-party databases (e.g. Dun & Bradstreet, Dow Jones, etc.) or official corporate registries to obtain accurate and verifiable information.
	Approach	Corporate registry search	
	Resources	https://opencorporates.com/	
If a registrant's address is available from Whois records, what can you learn from searches of that address? Do other entities appear to be located there?	Element / Feature	Domain registrant address	Note If a given domain has been registered using a privacy registrar (e.g. GoDaddy, Domains By Proxy), the IP and physical address typically resolve to where the privacy registrar is domiciled rather than the location of the domain's ultimate beneficial owner.
	Approach	Reverse address search	
	Resources	https://www.whitepages.com/reverse-address	

2. Website design			
Is this site custom made or has it been developed using a commercial template? Have you searched for other sites built using the same template?	Element / Feature	Website template	
	Approach	WordPress theme or template search	
	Resources	https://www.codeinwp.com/find-out-what-wordpress-theme-is-that/	
Has this site's content or design changed since it was registered? What did it look like previously?	Element / Feature	Historical versions of the website	
	Approach	Cached website search	
	Resources	https://archive.org/ https://cachedview.com/	

3. Website content			
Who are the authors of the content hosted on the target site and what can you learn about them? Have the articles posted on the website appeared elsewhere online?	Element / Feature	Website authors and/or snippets of article text	Note As with corporate registry data, while open source tools can be useful for identifying relevant media articles, it may be worth considering searches of paid news media databases such as Factiva, LexisNexis or ProQuest to ensure completeness.
	Approach	General and adverse media search	
	Resources	https://news.google.com/ https://newspapermap.com/	
Have the images or videos hosted on the target site appeared elsewhere online?	Element / Feature	Images, video or other multimedia content on the site	Note The effectiveness of reverse image search tools varies depending on location and image format. For example, Yandex is widely considered to be a leader in facial recognition while Google's scenery database is more comprehensive than most other reverse image tools.
	Approach	Reverse image search	
	Resources	https://www.osintcombine.com/reverse-image-analyzer https://tineye.com/ https://www.bellingcat.com/resources/how-tos/2019/12/26/guide-to-using-reverse-image-search-for-investigations/	

4. External links			
Who is linking to the target site elsewhere online?	Element / Feature	Domain URL	Note Checking backlinks to a target site can be an effective way of understanding the site's role in a broader network of disinformation sites.
	Approach	Backlinks check	
	Resources	https://ahrefs.com/backlink-checker https://www.thehoth.com/backlinks-checker/	

Going Deeper on Definitions

Computational propaganda, disinformation, misinformation, malinformation, covert information operations: the growing sector of research on the use of technology to spread false information or to mislead has already collected a wealth of terms for its object of study. This is an area that continues to suffer from overlapping and poorly delineated definitions of the problem, as well as the related tactics, techniques and strategies employed. This is not just a result of the comparative youth of this field of study or the clashing priorities and viewpoints of different researchers. It also speaks to the pace of change in possible activities enabled by an ever-evolving technological ecosystem.

There are a number of fields, outside research and academia, that have their own nomenclature for these types of threat. The legal sector's existing language around defamation and libel constitutes one narrow area of what might now be termed disinformation. Some states have drawn up legislation providing new definitions of disinformation, for example in France: 'Inexact allegations or imputations, or news that falsely report facts, with the aim of changing the sincerity of a vote' (Assemblée nationale, 2018). The buzzword of 2016 – 'fake news' – remains an oft-cited term in media reporting on the issue, although politicians seeking to undermine legitimate criticism frequently co-opt it. Each technology company has its own definitions of disinformation and misinformation, constantly in flux.

In a field as new and ever evolving as digital disinformation, vague or imprecise definitions can lead to problems not only in how research is received, used or cited by others, but also for the potential to open up institutions to criticism or potential legal challenge in some contexts. All organisations should prioritise the transparency and clarity of definitions in this area of work, even where those definitions do not remain stable over time, and are influenced by moveable threats.

Defining Disinformation Activities

One thing has become abundantly clear: the study of disinformation cannot focus solely on false information. False identities, false communities and false popularity are all part of the playbook of nation states and ideologically motivated actors. François' 'ABC' framework for disinformation speaks to the range of considerations required to conduct comprehensive disinformation research (François, 2019). The framework was recently supplemented by 'D' for distribution mechanisms for disinformation by Alaphilippe in order to include the role of platform products in amplifying, promoting or targeting disinformation (Alaphilippe, 2020).

Yet illicit actors, behaviour, content or distribution mechanisms are each contested terms, relying heavily on the perspective, values and context of the researcher. Here we discuss the definitions that have been most broadly accepted in the field and that encapsulate as wide a range of disinformation activities as possible, within a coherent framework. **There is value in using established and widely shared definitions of the problem in order to compare research approaches, data and findings across organisations and contexts.**

Many of the most widely used definitions still rest on an understanding of intent, which is in most circumstances problematic for researchers to understand or evidence. Recognising this, Spies at the MediaWell project (2020) adopts a broad terminology in discussing incidents of false information or misleading information, where they

> *recognize the limitations of an intention-based distinction between dis- and misinformation, and suggest that they be considered together in a way that allows for their mutability, referring to 'dis- and misinformation'.*

Where intent cannot be evidenced, such caveats are useful and it is important to communicate them.

Defining False Information

There are a number of reliable and widely accepted definitions of disinformation or misinformation that focus narrowly on the issue of false or misleading content. These include:

- Wardle and Derakhshan (2017): disinformation is 'information that is false and deliberately created to harm a person, social group, organization or country'.
- Wardle and Derakhshan (2017): misinformation is 'information that is false, but not created with the intention of causing harm'.
- MediaWell project (2020): disinformation is 'a rhetorical strategy that produces and disseminates false or misleading information in a deliberate effort to confuse, influence, harm, mobilize, or demobilize a target audience'.
- MediaWell project (2020): misinformation is 'false or misleading information, spread unintentionally, that tends to confuse, influence, harm, mobilize, or demobilize an audience'.
- Jack (2017): disinformation is 'deliberately false or misleading' information.

Defining Influence Operations

For researchers who focus on the digital field of disinformation, there are a number of established fields of study that shine light on the nature of influence communications writ large that should not be laid aside. They include the study of propaganda, which does not revolve around falsity but around the intention to persuade. These are some definitions that consider a broader range of information activities conducted by nation states and other actors, which are not necessarily deceptive or false and may bridge the online and offline:

- Benkler, Faris and Roberts (2018): propaganda is 'the intentional manipulation of beliefs' or 'communication designed to manipulate a target population by affecting its beliefs, attitudes, or preferences in order to obtain behavior compliant with political goals of the propagandist'.
- Jack (2017): information campaigns are 'organized communicative activities that aim to reach large groups of people. With many information campaigns, there is no question that they are deliberate attempts to persuade. The terms advertising, public relations, public diplomacy (or public affairs), information operations, and propaganda all describe deliberate, systematic information campaigns, usually conducted through mass media forms – the press, broadcast media, digital media, public events and exhibitions, and so on… Persuasive campaigns may involve accurate information, misinformation, disinformation, or a mix of all three.'

Understanding Disinformation Activities, Behaviours and Content

Since 2019, ISD has used 'malign information activities' as its framework for analysis, defined as 'activities that use online products, media systems or platforms with the outcome of deceiving audiences, distorting the available flow of information or conducting illegal activities'.

The above definition is specific to ISD's research objectives, which go beyond the study of false information. The definition intentionally includes online incidents of existing illegal harms, which can include, but is not limited to, harassment, hate speech or terrorist recruitment. It also incorporates activities from any type of state or non-state actor, in line with ISD's own remit and mission. Finally, the definition includes practices of 'distortion' of the information space, in order to incorporate overt or covert activities that affect the flow of information.

Other expert institutions have provided definitions featuring a range of activities and tactics used to deceive or mislead online, including the definition of 'malinformation' from First Draft (Wardle, 2017): 'factual information released to discredit or harm a person or institution, such as doxing, leaks, and certain kinds of hate speech'.

Alaphilippe, A. (2020) 'Adding a "D" to the ABC disinformation framework', Tech Stream, Brookings, 27 April 2020, https://www.brookings.edu/techstream/adding-a-d-to-the-abc-disinformation-framework/.

Assemblée Nationale (2018) 'Lutte contre la manipulation des informations', 22 March 2018, http://www.assemblee-nationale.fr/dyn/15/dossiers/alt/lutte_fausses_informations.

Benkler, Y., Faris, R. and Hal Roberts (2018) 'Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics', Oxford University Press, 29 November 2018, https://www.amazon.co.uk/Network-Propaganda-Manipulation-Disinformation-Radicalization/dp/0190923628.

François, C. (2019) 'Actors, behaviors, content: a disinformation ABC', Transatlantic Working Group, 20 September 2019, https://science.house.gov/imo/media/doc/Francois%20Addendum%20to%20Testimony%20-%20ABC_Framework_2019_Sept_2019.pdf.

Jack, C. (2017) 'Lexicon of Lies: Terms for Problematic Information', Data & Society, 2017, https://datasociety.net/pubs/oh/DataAndSociety_LexiconofLies.pdf.

Kwan, V. (2019) ***Responsible Reporting in an Age of Information Disorder***, First Draft, https://firstdraftnews.org/wp-content/uploads/2019/10/Responsible_Reporting_Digital_AW-1.pdf.

Spies, S. (2020), 'Defining "Disinformation"', 29 April 2020, MediaWell, https://mediawell.ssrc.org/literature-reviews/defining-disinformation/versions/1-1/.

Wardle, C. (2017) 'Fake news. It's complicated', First Draft, 16 February 2017, https://firstdraftnews.org/latest/fake-news-complicated/.

Wardle, C. and Hossein Derakhshan (2017) 'Information Disorder: Toward an interdisciplinary framework for research and policy making', 27 September 2017, https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

We are a global team of data analysts, researchers, innovators, policy-experts, practitioners and activists - powering solutions to extremism, hate and polarisation.

The Institute for Strategic Dialogue (ISD) is an independent nonprofit organisation dedicated to safeguarding human rights and reversing the rising global tide of hate, extremism and polarisation. We combine sector-leading expertise in global extremist movements with advanced digital analysis of disinformation and weaponised hate to deliver innovative, tailor-made policy and operational responses to these threats.

Over the past decade, we have watched hate groups and extremist movements deploy increasingly sophisticated international propaganda, influence and recruitment operations, skillfully leveraging digital technology, and often boosted by hostile state actors. Alongside an exponential spike in violence (conflict, hate crime, terrorism), societies around the world are being polarised. At ballot boxes, populists have made significant gains and authoritarian nationalism is on the rise. If left unchecked, these trends will existentially threaten open, free and cohesive civic culture, undermine democratic institutions and put our communities at risk. Progress on the major global challenges of our time – climate change, migration, equality, public health – threatens to be derailed.

We can and must turn the tide. Help us build the infrastructure to safeguard democracy and human rights in the digital age. We believe it is the task of every generation to challenge fascistic and totalitarian ideologies and to invest in reinforcing open, democratic, civic culture.

ISD draws on fifteen years of anthropological research, leading expertise in global extremist movements, state-of-the-art digital analysis and a track record of trust and delivery in over 30 countries around the world to:

1. Support central and local governments in designing and delivering evidence-based policies and programmes in response to hate, extremism, terrorism, polarisation and disinformation

2. Empower youth, practitioners and community influencers through innovative education, technology and communications programmes.

3. Advise governments and tech companies on policies and strategies to mitigate the online harms we face today and achieve a 'Good Web' that reflects our liberal democratic values

Only in collaboration with all of these groups can we hope to outcompete the extremist mobilization of our time and build safe, free and resilient societies for generations to come. All of ISD's programmes are delivered with the support of donations and grants. We have the data on what works. We now need your help to scale our efforts.

If we succeed in empowering just a small minority of the silent majority with the insights, knowledge and tools they need, we have won.

ISD | Institute for Strategic Dialogue

The Institute for Strategic Dialogue (ISD) is an independent nonprofit organisation dedicated to safeguarding human rights and reversing the rising global tide of hate, extremism and polarisation. We combine sector-leading expertise in global extremist movements with advanced digital analysis of disinformation and weaponised hate to deliver innovative, tailor-made policy and operational responses to these threats.

www.isdglobal.org

PO Box 75769 | London | SW1P 9ER | UK