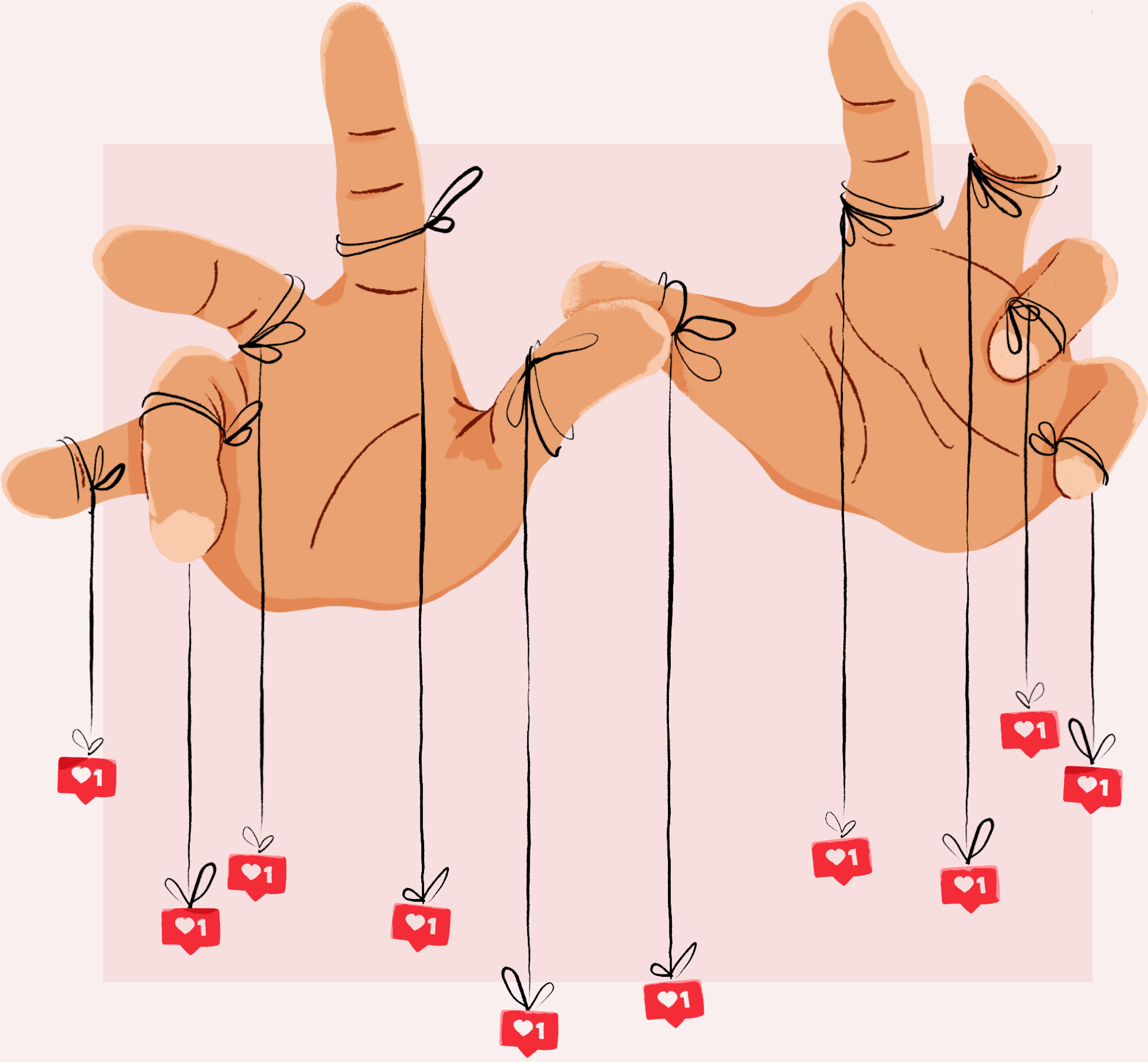
ISD | Institute for Strategic Dialogue

Toolkits → Systems → Coalitions

# Developing a Civil Society Response to Online Manipulation

Carl Miller
Chloe Colliver



Online Manipulation Resource Series

2/2

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.

Toolkits → Systems → Coalitions

# Developing a Civil Society Response to Online Manipulation

## Contents

# Vision

SUMMARY

This document presents a vision for a grand, civil societal response to online manipulation: developing the capability to detect it; the coalitions to confront it, the strategies to prevent it, and the structures of cooperation and funding that are needed to do both across all the causes and issues that it now threatens to undermine. It is a response that must combine the specialisation and sophistication that comes with scale with those particular strengths that civil society always has: its diversity, transparency, capacity to connect with marginalised voices, and its bedrock of humane and humanising values.

Research has gradually revealed the extent to which online manipulation has been weaponised to affect societies in almost every important way that society works: its politics and beliefs, its values and identities, the problems that society sees with itself, and the activism and mobilisations that thereby result as it tries to change itself. From the global reaches of geo-politics to the very local, from formal elections to struggles over culture, language and heritage, it has formed a kind of background hum to recent history.

Yet while almost everyone is touched in one way or another by online manipulation, only a tiny part of society has generally been involved in confronting it. It has largely been an invisible struggle: on the one side the practitioners of illicit campaigns, whose identities, interests and real agendas have mostly remained in the shadows; and on the other the defensive teams employed by the tech giants. Treated principally as a question of user experience and platform integrity by the platforms themselves, the counter-measures they have taken and their effectiveness have in most cases remained as mysterious as the illicit campaigns they seek to stop.

Between the practitioners of online manipulation and the platforms themselves sits civil society. Over the last five years, a growing collection of academic institutions, think tanks and advocacy organisations have become involved in highlighting instances of online manipulation and the debates about how to stop it. A loosely defined sector has emerged, bringing together actors who specialise in researching disinformation, targeted harassment, and weaponised hate, with those based in the traditions of human rights, consumer rights, privacy and corporate accountability. Transparent and innovative models for agile detection of online manipulation have already emerged from this sector as it stands, detecting and exposing platform manipulation around the globe.

This sector has faced a number of formidable and systemic challenges, however. Historically, it comprises reasonably small, agile organisations, often built around a particular cause, theme or function. Technology development often occurs in *ad hoc* circumstances, frequently based on event-specific, time-limited pockets of funding won in competition against other parts of civil society.

This document presents a vision for a grand, pan-civil societal response to online manipulation. In part, it argues, this will come down to capability: building a pooled detection capacity to function as a transparent, public interest alternative to those built by the tech giants. In part, it will require new organisational philosophies and forms of co-operation, and in part new approaches to funding and support. Overall, the vision tries to unite the sophistication and specialisation that a scaled response can confer, with everything that makes civil society a crucial part of the solution: its diversity, capacity to connect with marginalised voices and communities, transparency and passionate support for the values, causes and issues that its members, supports and workers believe in and that online manipulation itself now threatens to undermine. Tech platforms will con-

**THE VISION**

This document presents a vision for a grand, sectoral response to online manipulation. In part, it argues, this will come down to capability: building a pooled detection capacity to function as a transparent, public interest alternative to those built by the tech giants. In part, it will require new organisational philosophies and forms of co-operation, and in part new approaches to funding and support tinue to be limited both by business interests and by single-platform detection research. Government teams dedicated to identifying online manipulation will remain largely restricted to detecting instances of 'foreign' interference. Civil society is the only place where independent, cross-platform, comprehensive research on online manipulation can realistically take place in ways rooted by human experience and societal values. This document attempts to lay out what an ideal version of that work might look like.

### Approaches to Date

In 2019, ISD conducted an evaluation of work that had been undertaken to detect illicit influence online, especially influence operations targeting elections. We identified three broad types of approach:

**Bad Actor Driven Research:** Something would be found from observation of known bad actors – often a new behaviour, campaign, specific message and so on – which would trigger a wide investigation of social media and news sites to assess whether and how far this campaign had reached into the mainstream. This method could either be pursued manually, or by using social media analytics platforms. It tended to be most effective on platforms such as Twitter, where data could be collected based on accounts, rather than spaces or language. Platforms such as Facebook that offers far less visibility regarding the holistic activity of an account were less amenable to this method. Examples include the Hamilton 2.0 Dashboard↗ by Alliance for Securing Democracy or ISD's work monitoring information operations conducted by the global far-right↗ and conspiracy theory networks↗.

**Message Driven Research:** This began by identifying a message, theme or story, a 'share of voice', or heightened salience of an issue that it was considered important to learn more about. These would typically be identified using social media data analytics tools, and would then trigger an investigation to attempt to find the origin, intent, interests, groups or individuals behind it, usually using a blend of social media analytics and open source intelligence (OSINT). Examples include Bellingcat's investigation↗ into the disinformation campaign against the White Helmets in Syria or the joint investigation↗ between BBC, Australian Strategic Policy Institute and Bellingcat analysts about a co-ordinated social media campaign aimed at the Indonesian province of Papua using the #WestPapua↗ and #FreeWestPapua↗ slogans.

**Spotting Anomalies in Big Data:** This began with the broad appraisal of social media data, especially patterns around engagement and sharing. This might either be through appraisal using a social listening tool, or more bespoke methods composed by the analyst themselves. Patterns identified would then trigger more specified research endeavours, usually an OSINT investigation. Examples include the George Washington University's School of Media & Public Affairs' investigation↗ into pro-AfD election campaign activity on Facebook.

A key finding of the evaluation was that each of these approaches had distinct groups of strengths and weaknesses. Investigations based on immersive, manual observation of known and presumed bad actors powerfully leveraged the existing subject matter expertise of researchers, but tended to reflect what the researcher already knew, rather than discovering something new. Message driven research was an approach stronger on discovery, but tended to miss associations between how different (and sometimes ideologically conflicting) narratives could combine into coherent influence campaigns that could expose how influence campaigns worked. Anomaly detection linked discovery to attribution effectively in some cases, but was extremely resource intensive and usually platform specific.

The next step towards a civil society detection capability is to harness the complementarity of these different approaches. The idea is to use the strengths that each approach confers to mitigate the weaknesses of the others. Overall the aim is to construct a workflow that moves flexibly across these approaches, borrowing from each at different times.

**CURRENT APPROACHES**

- Bad Actor Driven Research: Something would be found from observation of known bad actors – often a new behaviour, campaign, specific message and so on – which would trigger a wide investigation of social media and news sites to assess whether and how far this campaign had reached into the mainstream.
- Message Driven Research: This began by identifying a message, theme or story, a 'share of voice', or heightened salience of an issue that it was considered important to learn more about.
- Spotting Anomalies in Big Data: This began with the broad appraisal of social media data, especially patterns around engagement and sharing. This might either be through appraisal using a social listening tool, or more bespoke methods composed by the analyst themselves.

# Key Design Principles

The eight core design principles that must be at the heart of a civil society response

An ideal capability for civil society should not be conceived as a toolkit. There are lots of reasons why this is the wrong way of thinking about it: most of the tools that are currently needed do not yet exist, and the tools that do exist broadly do not do what is needed. Having a toolkit implies that the problem of detecting online manipulation can be solved by equipping analysts with a series of over-the-counter services that they can variously pick up and use as circumstances demand. This is not, however, really the case. The problem of detection is much more likely to be surmounted by something that should be thought of as an evolving, deployable and highly customisable system.

This document lays out what such a system for civil society should look like. It is in part research strategy, in part technology architecture, in part human capabilities, and in part collective memory. Across all of these areas, it is best defined by a series of design principles that we propose must be at the heart of this undertaking.

## From Society, For Society 1/8

## Be Data Hungry 2/8

## Shareable Modules 3/8

## Fast Tech Dev 4/8

## Reactive Data Vis 5/8

## Learn from Itself 6/8

## From Evidence to Ideas 7/8

## Human Impact at its Heart 8/8

## Key Design Principles

This section explains each of the design principles in greater detail, looking at the ways that they might be achieved, and the various people, skills and technologies that must be brought together in order to do so.

### 1/8 From Society, For Society

The system must be plugged into civil society in a number of ways: its priorities and direction should be informed by those of many groups and communities across civil society; it should work in ways that are transparent and understandable to civil society, and it should produce outputs that allow for a civil societal response, that are actionable.

### 2/8 Be Data Hungry

It must leverage the full opportunities available for civil society researchers to acquire data from all the platforms and online spaces relevant to illicit online manipulation, and this will typically stretch across a wider swathe of the internet than is often researched.

### 3/8 Shareable Modules

It should then have a detection capability to identify and filter social media data according to whether it conforms to one of a series of behaviours that relate to illicit online influence. This detection capability should be sensitive to platform, but operate across platforms. It is understood that illicit influence online frequently occurs across a number of platforms, often functionally separated for the purposes of planning, co-ordination and execution. Detections made on one platform may present either data collection opportunities on another platform, or input into the detection methodology on another platform.

### 4/8 Fast Tech Dev

The system will never be in a settled state. It must have a reactive technology development capacity where a team of developers is tasked to add additional technological capability, driven by analytical teams who are face-to-face with the data that is being analysed. This assumes that a number of important technology requirements will only be discovered through the continued and practical use of the system.

### 5/8 Reactive Data Vis

It should have a visualisation and analysis function wherever the machine-driven parts of the detection system have a user-friendly touchpoint with humans. This includes data interrogation visualisations at stages in the system where analysts must give manual guidance or make interventions to the machine, and also a visualisation surface to produce results.

### 6/8 Learn from Itself

It has a cyclical discovery function. The system must help analysts find examples of online manipulation which are not known, on the basis of what is known. The workflows and technologies used to conduct illicit influence (and so the detection opportunities they create) will change, as will the issue areas, messages and narratives they seek to interact with, and the underlying social dynamics and processes they seek to either exploit or influence. As far as it is possible, the system must be designed cyclically, such that its outputs can be used as further inputs. This means that over time the system will learn from itself, and so evolve as the phenomena that it tracks themselves change in nature and scope.

### 7/8 From Evidence to Ideas

The empirical outputs of the system contribute to, and draw from, conceptual and definitional work. This is an area which continues to suffer from overlapping and poorly delineated definitions of the problem phenomenon itself, and the tactics, techniques and strategies related to it. Any detection system must also produce a systematic and continuous effort to develop the abstracted concepts able to organise and understand online manipulation. A feedback loop is therefore needed to link the evidence generated, and the ideas that organise and make sense of that evidence.

### 8/8 Human Impact at its Heart

The system must incorporate and apply methodologies drawn from psychology and sociology to understand the scale and nature of real-world harm that is produced from online manipulation. This understanding can therefore help to prioritise and triage the detection of the behaviours that cause different kinds of harm.

# From Society, For Society

A civil society response must…

…be plugged into civil society in a number of ways: its priorities and direction should be informed by those of many groups and communities across civil society; it should work in ways that are transparent and understandable to civil society, and it should produce outputs that allow for a civil societal response, that are actionable.

A civil society capability for detecting online manipulation should not try simply to reproduce the capabilities of the tech giants. It can take a different approach to the problem by engaging with similar underlying technologies, but do so in ways that harness the signature strengths of civil society:

### An 'open agenda'

The research agenda and priorities – and indeed the underlying idea of harm that the system is designed to detect – can be set and reset by a much wider array of voices from across society than any analogous system within big tech. Groups that are engaged with the victims of influence operations, those vulnerable to them, and the issue areas frequently targeted by influence operations are all key partners to engage. The system's research agenda can be set by these groups through the direct submission of either the examples or effects of influence operations. This can happen either through informal contact with the detection team, or through a more formalised ticketing and feedback system.

### Transparency

The system itself can and should be made much more transparent to members of civil society. Of course, the efficacy of detection systems can sometimes be undermined by transparency, especially if a system allows bad actors to reverse engineer and then game detection methodologies. However, using combinations of trusted partners and credentialed access, this risk can be mitigated to turn what are usually black box systems into transparent examples of civil technology, including:

- ↗ open sourced models
- ↗ open source data management software and workflows
- ↗ models which are trained by a wide variety of different actors
- ↗ detailed data outputs which can be consumed and analysed by a wide variety of actors
- ↗ outputs which are publicly visible, available and capable of contributing to public and policy debate.

### Networked insights

The system can be designed to co-operate with larger and looser networks as well as smaller, more dedicated teams. One of the key opportunities to do this is in the open source investigation of possible online manipulation campaigns (see discussion in the second part of this document). Leveraging volunteer-based networks would allow a civil societal response to increase radically the languages it could operate across, the skills it might tap, and the impacts it might have.

# Be Data Hungry

A civil society response must…

…leverage the full opportunities available for civil society researchers to acquire data from all the platforms and online spaces relevant to illicit online manipulation, and this will typically stretch across a wider swathe of the internet than is often researched.

Data can be collected by civil society researchers in a number of ways: through direct application programming interface (API) access to platforms, by 'scraping' forums and websites, and by integrating with a number of third-party data providers who offer relevant datasets, usually as a commercial enterprise. 'Crowdsourced' data streams, such as the Infotagion↗ reporting platform enable additional access beyond platform-regulated data streams.

This landscape of data availability is highly varied and constantly changing. The primary platforms themselves usually offer a number of APIs, each of which offer different kinds of data depending on the kind of query. Some platforms have released additional data to selective groups of academic researchers, including the Social Science One and the Social Science Research Council  grants or Twitter's call for research on the platform's 'conversational health'.

The structure of data collected from various platforms differs: a system should attempt to find as much commonality in the endpoints provided by platform APIs or other data streams in order to enable automated cross-platform analysis as well as manual cross-platform analysis. This will also require conceptual work (see Principle Seven) to determine what types of spaces, behaviours or metrics can fairly be compared across platforms and which cannot.

Crucial across any data collection strategy are the queries that are made to APIs or data repositories to return information. Common approaches to queries in this area of work include handcrafted lists of:

**Keywords**
Selected on the basis of apparent relevance to the topic in question

**Accounts, channels or pages**
Selected on the basis of apparent relevance to a topic in question or previous relevant behaviour

**Domains or links**
Selected on the basis of apparent relevance to the topic or actors in question.

The system must move beyond using handcrafted lists for data collection. It is outside the scope of this document to explain this at length, but methodologies can be constructed that combine subject matter expertise with text analytics to identify (and retire) keywords systematically. This will be a data-driven way to identify queries that are relevant to what the system must detect – search terms, account names, links, hashtags and so on.

A civil society response must...

...have a detection capability to identify and filter social media data according to whether it conforms to one of a series of behaviours that relate to illicit online influence. This detection capability should be sensitive to platform, but operate across platforms. It is understood that illicit influence online frequently occurs across a number of platforms, often functionally separated for the purposes of planning, co-ordination and execution. Detections made on one platform may present either data collection opportunities on another platform, or input into the detection methodology on another platform.

This system primarily focuses on identifying online phenomena that tend to be deliberately hidden and is thus different from most forms of social research. Detection methodologies are, in principle, descriptions of ways that manipulation behaviour online differs from all other forms of behaviour online. These descriptions might be models, rules, or based on an analyst's judgement.

We envisage detection capabilities acting on three different levels:

**Content level:**
Forms of online manipulation that can be detected by the nature of the content itself. Content-level detection can be used to find the propagation of explicit falsehoods, conspiracy theories, the existence of voter-suppression or voter-depression information, and content which calls for action against minority groups, for instance.

**Account level:**
Forms of detection which are only possible once the individual behaviour of a particular account is summarised and aggregated. For example, account-level detection can identify fully or semi-automated accounts, compromised accounts, and forms of influence 'activation' where an account's behaviour suddenly shifts in unusual ways.

**Network level:**
Detection of online manipulation only visible once the behaviour of an entire network is appraised and summarised. This is necessary to identify false amplification networks or 'bot nets' for networks engaged in 'covert inauthentic behaviour', to identify links between accounts, pages, channels and domains, or co-ordinated harassment campaigns.

The techniques of online manipulation, and methods to hide them, change constantly. Detection therefore forms part of an unceasing dynamic of action–reaction by actors on both sides, resulting in continual tactical and technological evolution.

It is for this reason that a civil society detection system must have a modular and composite structure, whereby detection methods of different kinds can be built by different teams, with different specialisms and at different times. These can be pooled into a central library, and deployed in different configurations as circumstances demand.

A civil society response must...

...never be in a settled state. It must have a reactive technology development capacity where a team of developers is tasked to add additional technological capability, driven by analytical teams who are face-to-face with the data that is being analysed. This assumes that a number of important technology requirements will only be discovered through the continued and practical use of the system.

The fundamental challenge any system design, research ambition or toolkit description faces is that it must detect a constantly changeable universe of behaviours, identifiable by a changeable universe of signals. Many new challenges will only be discovered as others are overcome, and future forms of online manipulation cannot be anticipated yet. The need for new technology will only become apparent as the system is deployed and used.

Therefore, there is a need for an intense reciprocal dynamic between toolmakers and data analysts. Analysts who detect online manipulation must be able to identify gaps and limitations in the tools that they have, and be able to work with developers who are capable of changing those tools. Tools and an analytical approach can therefore evolve side by side.

A civil society response must...

...have a visualisation and analysis function wherever the machine-driven parts of the detection system have a user-friendly touchpoint with humans. This includes data interrogation visualisations at stages in the system where analysts must give manual guidance or make interventions to the machine, and also a visualisation surface to produce results.

The system must have two different kinds of visualisation capabilities:

### A 'vis-analytic' functionality

A 'vis-analytic' functionality within the system to facilitate the interaction between machine and analyst. The system must unite the subject matter expertise of humans with the capacity of machines to operate across great scales of data, so analysts will have to calibrate, check, audit, re-define and re-direct how the system operates at many different touch-points within it. There is therefore a need to have a data visualisation capacity at those touchpoints, to allow analysts to understand the results not only at the end of the process, but also at key sub-analytical stages within it. This capacity must be reactive, in the sense that the visualisations must be built around helping the analyst make the key identifications in the data necessary for any particular step.

### A visualisation front-end

A visualisation front-end to allow journalists, policymakers, other researchers and the general public to understand the detections that the system has made quickly and intuitively. It must also capture the limitations and uncertainty present in the results, in order to communicate them in an honest way. Many of the underlying technologies used for detection produce results which are inherently probabilistic in nature, and this presents a significant challenge for end-point visualisations to capture and display as important caveats to the findings.

A civil society response must...

...have a cyclical discovery function. The system must help analysts find examples of online manipulation which are not known, on the basis of what is known. The workflows and technologies used to conduct illicit influence (and so the detection opportunities they create) will change, as will the issue areas, messages and narratives they seek to interact with, and the underlying social dynamics and processes they seek to either exploit or influence. As far as it is possible, the system must be designed cyclically, such that its outputs can be used as further inputs. This means that over time the system will learn from itself, and so evolve as the phenomena that it tracks themselves change in nature and scope.

Journalists and civil society tend to be able to identify individual examples and instances of online manipulation. A key capability gap, however, is in being able to move from anecdotal examples to a comprehensive and exhaustive mapping. The system must therefore allow analysts to discover additional new parts of the phenomenon on the basis of what is already known about it. This is especially important with regard to evolving phenomena. The system must have a 'cyclical' design, such that it can learn from itself.

This discovery function largely relates to turning detections that the system makes into new data collection inputs:

### New data collections

Relevant detections the system outputs can be used as the basis for new data collection inputs into the system. This will mostly be either new accounts, channels, groups, links, hashtags or subreddits that are identified as relevant and can be collected, or new keywords and phrases that can be used in data collections. As mentioned above, this can be especially powerful when applied across a number of platforms: for instance, data is collected from YouTube channels on the basis of links shared on Twitter, and Facebook groups are identified on the basis of posts made on Reddit.

### New detections

It is also possible that the outputs of some detections may be used as inputs for others. For instance, the language used by a group of known bad actors might be turned into a semantic model used for a new detection. Or the follower activity of a suspicious network on Twitter might be used as an indicator of an illicit propaganda network on Facebook.

A civil society response must...

...have the empirical outputs of the system contribute to, and draw from, conceptual and definitional work. This is an area which continues to suffer from overlapping and poorly delineated definitions of the problem phenomenon itself, and the tactics, techniques and strategies related to it. Any detection system must also produce a systematic and continuous effort to develop the abstracted concepts able to organise and understand online manipulation. A feedback loop is therefore needed to link the evidence generated, and the ideas that organise and make sense of that evidence.

Conceptual development regards the importance of developing not only an empirical and descriptive understanding of online manipulation, but also the ideas, definitions and concepts that can make sense of, and organise, the empirical outputs. The principle applies in a number of disciplines; for example, information security analysts have developed methodologies for understanding threat actors, and systematising their workflows and capabilities.

It is vitally important that empirical and conceptual work feed off and inform each other. Simply developing an empirically richer picture of online manipulation alone will not by itself reveal the many ways that it can be responded to, especially through changing the incentives, costs, risks and opportunities available to the actors who currently seek to undertake such manipulation.

Likewise, empirical work can help inform abstract thinking about what harms online really look like. Liberal philosophy underlies many of our ideas around what 'licit' and 'illicit' forms of influence really are. Geo-political experts can help determine the true motivation that may drive online manipulation. Democracy theory can help us understand new forms of electoral interference.

A civil society response must...

... incorporate and apply methodologies drawn from psychology and sociology to understand the scale and nature of real-world harm that is produced from online manipulation. This understanding can therefore help to prioritise and triage the detection of the behaviours that cause different kinds of harm.

Civil society's response to online manipulation is predicated on the idea that there are activities online that can cause a number of broad harms to the democratic, political and social lives of its targets. However, a detection system which is based purely on online information is often unable to measure the wider effects that this online information causes.

Understanding of individual and societal harm requires the application of methodologies drawn from psychology and sociology, and the connection of online datasets with other ways of understanding belief and behaviour. There is a need for analysts from a parallel research effort to conduct this work, in order to understand the scale and nature of harm that is produced, and therefore how to prioritise and triage the detection of the behaviours that cause them.

Detecting the impact of online manipulation could include the use of polling and other attitudinal data, for example constructing an 'impact panel' in communities and countries that are targeted by online manipulation. It could comprise two kinds of cohort: a panel representative of the country, and another representative of more specific 'at-risk' groups. The identification of at-risk groups can be based on those that previous information warfare campaigns have targeted, including older demographic groups, people who have never voted before, and people who already hold conspiracy theoretical or radical political beliefs. The impact panel should be polled regularly to look for attitudinal and behavioural indicators of possible influence operations: from overt awareness and belief in conspiracy theories, to levels of trust in government, to background emotional factors – anxiety, distress, fear, outrage and so on.

These kinds of research are especially valuable because they can begin to allow researchers to isolate the exposure to illicit influence campaigns from the wider information ecosystems which people live within. This would crucially contextualise the manner and scale of influence that is being illicitly exerted.

# Teams, Skills and Partnerships

The human infrastructure needed to enable the system to deliver results

The first part of this document described the key design principles we believe necessary in any capability for civil society that can detect online manipulation effectively. In the second, we describe the wider ‘human infrastructure’ needed to enable the system to effectively deliver results.

Teams and Skills 1/5

Mass Mobilisation: From Teams to Networks 2/5

Building the Sector: From Organisations to Coalitions 3/5

Building a Memory 4/5

High Risk, High Reward Technology Development 5/5

This section explains, in greater detail, each of the elements of the 'human infrastructure' that the system requires to effectively deliver results.

### 1/5 Teams and Skills

A civil society response needs to bring a number of skills and specialisms together. Clearly no single set of skills is sufficient to confront online manipulation. Data science alone is unlikely to uncover motivation, interests or identities. Manual analysts cannot cope with the sheer scale of social media data with which they are confronted. Any capability will only be successful if it interlocks a number of skills and specialisms together into a coherent workflow.

### 2/5 Mass Mobilisation: From Teams to Networks

A civil society response needs to take advantage of networks. As mentioned above, civil society's attempts to detect online manipulation enjoy a number of advantages over their commercial analogues. A key advantage is that they can leverage volunteer and membership-based networks to do things not possible by smaller teams.

### 3/5 Building the Sector: From Organisations to Coalitions

A civil society response needs to forge effective and meaningful coalitions. Much of this document is dedicated to describing a civil societal response to online influence as, necessarily, a combination of specialisms, skills, technology and know-how across a range of organisations. Beyond teams themselves, and even networks, coalitions should be forged across civil society to ensure that the use of such a capability, its emphasis and outputs are both effective and meaningful, and also reflect the priorities of the broadest possible base of partners and collaborators.

### 4/5 Longer-Term Development: Building a Memory

A civil society response needs to develop a memory that lasts longer than a single election cycle. Any detection system is usually deployed with a specific aim in mind: to protect an election, a community, a sector of work coming under attack, or to mitigate the fallout from a crisis or major event. As this system is deployed in reaction to these areas, it should be able to develop a longer-term memory of all the detections it makes, and how the problem of online manipulation and the task of detecting it has evolved.

### 5/5 High Risk, High Reward Technology Development

A civil society response needs to enable risk taking when developing detection technology. Within any particular deployment of the system, analysts and technologists will be under both time and financial pressure. They will have to make decisions regarding the technologies that are used, and how they are implemented, which will tend to prioritise safer, tested and better understood approaches over riskier and newer alternatives.

To work together effectively, we must...

...bring a number of skills and specialisms together. Clearly no single set of skills is sufficient to confront online manipulation. Data science alone is unlikely to uncover motivation, interests or identities. Manual analysts cannot cope with the sheer scale of social media data with which they are confronted. Any capability will only be successful if it interlocks a number of skills and specialisms together into a coherent workflow.

These are the type of workers required:

**Analysts and interrogators...**

who work closest to the data and are most familiar with how the system and its various components works, and have a series of generic skills regarding the analysis and management of large online datasets. This team configures and applies detections within the wider architecture.

**Reactive visualisation, technology and tool developers...**

who know about generic software development, especially related to data science, modelling, natural language processing, network analytics and other forms of machine learning. They react to novel analytical challenges and opportunities raised by analyst or interrogator teams, and apply backend software architectural development in response to these challenges.

**Data journalists and subject matter experts...**

who have the deepest level of subject matter expertise, which spans the actors who are seeking to conduct online manipulation, the targets of online manipulation, and the issue areas involved. They are able to understand the analytical outputs from the system, and identify anomalies, patterns, contrasts and consistencies with what is already known in order to identify the most promising leads for further investigation. They are a vital crossover link, both determining what leaves the system for further investigation, and also re-tasking the system with new thematic emphases.

**OSINT practitioners...**

who conduct targeted investigations of the most harmful, urgent and important detection that the system has made. They take leads from the system, and use a separate suite of OSINT tools to uncover the possible identities, motivations, ownership structures and hidden associations between online manipulation campaigns.

To work together effectively, we must...

...take advantage of networks. As mentioned above, civil society's attempts to detect online manipulation enjoy a number of advantages over their commercial analogues. A key advantage is that they can leverage volunteer and membership-based networks to do things not possible by smaller teams.

A number of functional parts of the system can be performed not just by teams, but by wider networks as well. These can help to determine the priorities of the system itself; to build it, to ingest it and to use the results. In addition to the points already made above, there are two opportunities here we believe to be worth foregrounding:

**Networks of technologists who contribute to modular detection**

The system must detect a number of online manipulation tactics, across a number of platforms, in combinations which are constantly evolving, and the methodologies possibly used to make these detections constantly change and spread across a number of specialisms. In the face of this challenge, it would be possible for networks of researchers to be leveraged, each working on a different detection, but contributing progress and outputs to a central system, itself capable of allowing new detection methodologies to be added in a modular way (see above).

**Networks of OSINT researchers**

Organisations such as Bellingcat have demonstrated enormous success by creating large, loose networks of OSINT practitioners who form organic collaborations around particular investigations. Something similar can be achieved with online manipulation, where the outputs of the detection system become the initial leads for the OSINT network.

The use of wider networks raises a number of issues. They:

- introduce the requirement for a community manager to co-ordinate between full-time staff and a wider network
- create additional risks to the operational security of the detection, and could make it more likely for detection methodologies to become known by adversaries
- create the need for additional communications and professional policy around how the network should operate, who can join it, what the expectations of them are, and so on.

These sit outside the scope of this paper to examine in greater detail, beyond the observation that the benefits of creating and using networks in this way may offset the additional costs and risks involved in doing so.

To work together effectively, we must...

... forge effective and meaningful coalitions. Much of this document is dedicated to describing a civil societal response to online influence as, necessarily, a combination of specialisms, skills, technology and know-how across a range of organisations. Beyond teams themselves, and even networks, coalitions should be forged across civil society to ensure that the use of such a capability, its emphasis and outputs are both effective and meaningful, and also reflect the priorities of the broadest possible base of partners and collaborators.

Civil societal coalitions should be structured for the following roles:

1. to determine overall direction
2. to react and mitigate
3. to build broader outputs and impacts
4. to build cross-organisational coalitions

### 1. To Determine Overall Direction

**Ensure the system reflects lived experiences**

The coalition should either include the groups and communities that are targeted by influence operations, or civil society organisations that work with these communities and groups. This is important both to create data inflows in the system, but also more broadly to ensure that the system and its deployment reflects the lived experiences of the communities that it should serve.

**Harness proactive threat intelligence**

As well as listening to the victims of influence operations, the coalition should also be able to research the actors who may be conducting it. Threat intelligence methods and skills typically sit outside the organisations involved in detecting influence operations, but could greatly inform the issue areas that detection efforts are deployed to protect, and the methods and behaviours they may exhibit.

### 2. To React and Mitigate

A coalition can combine different research teams who are staffed, skilled and briefed to respond to online manipulation at very different timescales:

- **A rapid response notification team** to react as quickly as possible to important changes or shifts in online activity. Staff would be equipped to verify the phenomenon as quickly as possible, and create near real-time updating outputs for members of the coalition and other stakeholders.
- **A community-facing mitigation team** comprising either one or a number of teams dedicated to working with the communities and groups targeted by illicit influence operations. With less emphasis on online research and analysis, this group would concentrate on producing mitigation toolkits, advertising campaigns, rapid training, aftercare provision and so on.
- **An in-depth investigation team** would constitute teams of blended forensic data science and OSINT investigators equipped for longer, more complex investigations and with a greater emphasis on attribution, discovery of additional facets of the phenomenon and more sophisticated forms of detection.
- **An impact assessment team**, as polling organisations that are capable of conducting quantitative attitudinal research, and social research organisations whose staff conduct interviews, ethnographic work and other forms of more qualitative research, are needed to understand what the human impacts of influence operations really are.
- **Subject matter and regional expertise**, with networks of area and geographic specialists able to adapt the system to the specificities of an individual threat vector. This would enable much more flexible engagements with in-country partners for instance around elections, or for wider cross-border networks to be supported on issue sets like climate change or public health.

### 3. To Build Broader Outputs and Impacts

The creation of coalitions opens the possibility of creating more specialised forms of output and output avenues. These can include:

- **An ‘immediate threat’ team** that has worked to create direct reporting channels into law enforcement agencies and groups under immediate threat, possibly across a range of different jurisdictions and with different responsibilities. The creation and maintenance of these relationships can be time consuming and ‘single points of contact’ can be created within this team with specific responsibilities in this area.

- **A strategic policy team** can use the range of empirical outputs that the coalition produces to create a broader picture of the scale of online influence, the venues within which it occurs, trends around the efficacy of platform enforcement, and so on. This team can also develop the relationships required to ensure that these outputs can reach and influence decision-makers within political institutions.

- **A civil society outreach and education team** can provide regular updates and organise events for civil society groups on trends, actors, networks, narratives and emerging threats, both as immediate and urgent reactions to new threats, and to longer-term, more strategic forms of resilience. These can include webinars, regular reports, memos and newsletters, published and non-public.

- **A combination of constant and high visibility media pressure**, as many different parts of the coalition will produce outputs that are newsworthy. These can take a number of forms, two of which are likely to be especially important:

  - a 'drumbeat' of media outputs to maintain consistent pressure highlighting failures of platform enforcement against online manipulation, remaining policy gaps, new threats and other developments requiring immediate response

  - longer-form, possibly collaborative, investigations to highlight the depth of the problem and exert pressure to remove key nodes of manipulation networks or push for systemic policy changes based on in-depth research into specific problem areas.

### 4. To Build Cross-Organisational Coalitions

Coalitions offer civil society the opportunity to specialise and scale their counter-influence efforts. In order to do so, however, they face a challenge that tech giants and other large corporate actors do not: they must find ways of working coherently outside established organisational hierarchies, clear lines of decision-making or even necessarily shared language, culture or workflows.

These are some important factors to consider when creating commonalities between organisations to allow coalitions to operate effectively:

- shared research and investigative resources, including (as stated above) platform lists; keyword lists; channel lists; shared definitions (where feasible); and shared tools, methods and models

- shared ethical frameworks between organisations to allow information to pass more freely between them; organisations need to have confidence in how data has been collected and how possible harm has been managed so they can receive and use it

- shared principles around communications and media to include shared understandings between coalition members about when and how to report on influence operations and the ways in which disclosures can themselves do harm, undermine faith in democratic processes or unrealistically elevate the likely influence of the detected operation itself.

There are also a number of structural aspects of civil society that coalition building must address. Civil society organisations often compete with each other for funding, and by extension for press coverage and public and political recognition for their work. A number of interventions can be made to align organisational interests to reward collaborative rather than competitive activity that civil society organisations conduct, for example:

- co-operative funding models initiated within philanthropy, which allow and reward the creation of outputs, as described above, that can increase detection capabilities across organisations, issue areas and events

- the creation of 'coalition-level governance', entailing the formalisation of decision-making and accountability structures which span across the different coalition members, including a shared decision-making apparatus, advisory board, oversight panel and so on;

- agreements to promote the mutual amplification of outputs for a multiplier effect and in order to mitigate some of the competition for public exposure;

- trust-building arrangements between organisations at more informal, cultural and intellectual levels, including reciprocal placements for researchers in partner organisations, shared fellowship programmes, combined online workplaces and mutually arranged events, think-ins, brainstorming sessions and away days.

## Longer-Term Development: Building a Memory

To work together effectively, we must...

... develop a memory that last longer than a single election cycle. Any detection system is usually deployed with a specific aim in mind: to protect an election, a community, a sector of work coming under attack, or to mitigate the fallout from a crisis or major event. As this system is deployed in reaction to these areas, it should be able to develop a longer-term memory of all the detections it makes, and how the problem of online manipulation and the task of detecting it has evolved.

It can do this by building:

**A reservoir of examples** of all detections of online manipulation, across all languages, platforms, cases, topics and themes, held in a format amendable to secondary investigation, and available to a wide number of researchers to draw out trends and higher-level insights.

**A library of detection approaches** and continuous internal evaluation and diagnostics, as over time it is likely that many detection approaches will be tried, with varying degrees of success. All evaluation material from these different forms of detection should be captured to allow for peer evaluation in the sector and to feed into future efforts.

## High Risk, High Reward Technology Development

To work together effectively, we must...

...enable risk taking when developing detection technology. Within any particular deployment of the system, analysts and technologists will be under both time and financial pressure. They will have to make decisions regarding the technologies that are used, and how they are implemented, which will tend to prioritise safer, tested and better understood approaches over riskier and newer alternatives.

These circumstances will militate against the rapid evolutionary development of civil society detection technology through the narrowing of alternatives used. There is therefore the requirement for a parallel effort that enables higher risk technology development pathways that may confer major new capabilities if they prove to be successful. These technologies may include:

**New forms of data collection**

Such as '360 degree' comprehensive browsing data from the consensual (and usually incentivised) use of browser plugins to provide researchers with full coverage of a person's online experiences and interactions in ways that are suitably anonymised and ethically handled

**The application of new forms of modelling**

Such as the deep-learning approaches associated with BERT ('Bidirectional Encoder Representations from Transformers'), which require significantly greater quantities of training data and experimentation that those typically required by in-use machine learning approaches

**Dynamic, real-time counter-messaging using targeted advertising**

Whereby detections immediately inform a series of responsive advertising campaigns using targeting criteria derived from the initial detections.

## Summary of key design principles

### 1/8 From Society, For Society

The system must be plugged into civil society in a number of ways: its priorities and direction should be informed by those of many groups and communities across civil society; it should work in ways that are transparent and understandable to civil society, and it should produce outputs that allow for a civil societal response, that are actionable.

### 2/8 Be Data Hungry

It must leverage the full opportunities available for civil society researchers to acquire data from all the platforms and online spaces relevant to illicit online manipulation, and this will typically stretch across a wider swathe of the internet than is often researched.

### 3/8 Shareable Modules

It should then have a detection capability to identify and filter social media data according to whether it conforms to one of a series of behaviours that relate to illicit online influence. This detection capability should be sensitive to platform, but operate across platforms. It is understood that illicit influence online frequently occurs across a number of platforms, often functionally separated for the purposes of planning, co-ordination and execution. Detections made on one platform may present either data collection opportunities on another platform, or input into the detection methodology on another platform.

### 4/8 Fast Tech Dev

The system will never be in a settled state. It must have a reactive technology development capacity where a team of developers is tasked to add additional technological capability, driven by analytical teams who are face-to-face with the data that is being analysed. This assumes that a number of important technology requirements will only be discovered through the continued and practical use of the system.

### 5/8 Reactive Data Vis

It should have a visualisation and analysis function wherever the machine-driven parts of the detection system have a user-friendly touchpoint with humans. This includes data interrogation visualisations at stages in the system where analysts must give manual guidance or make interventions to the machine, and also a visualisation surface to produce results.

### 6/8 Learn from Itself

It has a cyclical discovery function. The system must help analysts find examples of online manipulation which are not known, on the basis of what is known. The workflows and technologies used to conduct illicit influence (and so the detection opportunities they create) will change, as will the issue areas, messages and narratives they seek to interact with, and the underlying social dynamics and processes they seek to either exploit or influence. As far as it is possible, the system must be designed cyclically, such that its outputs can be used as further inputs. This means that over time the system will learn from itself, and so evolve as the phenomena that it tracks themselves change in nature and scope.

### 7/8 From Evidence to Ideas

The empirical outputs of the system contribute to, and draw from, conceptual and definitional work. This is an area which continues to suffer from overlapping and poorly delineated definitions of the problem phenomenon itself, and the tactics, techniques and strategies related to it. Any detection system must also produce a systematic and continuous effort to develop the abstracted concepts able to organise and understand online manipulation. A feedback loop is therefore needed to link the evidence generated, and the ideas that organise and make sense of that evidence.

### 8/8 Human Impact at its Heart

The system must incorporate and apply methodologies drawn from psychology and sociology to understand the scale and nature of real-world harm that is produced from online manipulation. This understanding can therefore help to prioritise and triage the detection of the behaviours that cause different kinds of harm.

## Summary of team, skills and partnership requirements

### 1/5 Teams and Skills

A civil society response needs to bring a number of skills and specialisms together. Clearly no single set of skills is sufficient to confront online manipulation. Data science alone is unlikely to uncover motivation, interests or identities. Manual analysts cannot cope with the sheer scale of social media data with which they are confronted. Any capability will only be successful if it interlocks a number of skills and specialisms together into a coherent workflow.

### 2/5 Mass Mobilisation: From Teams to Networks

A civil society response needs to take advantage of networks. As mentioned above, civil society's attempts to detect online manipulation enjoy a number of advantages over their commercial analogues. A key advantage is that they can leverage volunteer and membership-based networks to do things not possible by smaller teams.

### 3/5 Building the Sector: From Organisations to Coalitions

A civil society response needs to forge effective and meaningful coalitions. Much of this document is dedicated to describing a civil societal response to online influence as, necessarily, a combination of specialisms, skills, technology and know-how across a range of organisations. Beyond teams themselves, and even networks, coalitions should be forged across civil society to ensure that the use of such a capability, its emphasis and outputs are both effective and meaningful, and also reflect the priorities of the broadest possible base of partners and collaborators.

### 4/5 Longer-Term Development: Building a Memory

A civil society response needs to develop a memory that lasts longer than a single election cycle. Any detection system is usually deployed with a specific aim in mind: to protect an election, a community, a sector of work coming under attack, or to mitigate the fallout from a crisis or major event. As this system is deployed in reaction to these areas, it should be able to develop a longer-term memory of all the detections it makes, and how the problem of online manipulation and the task of detecting it has evolved.

### 5/5 High Risk, High Reward Technology Development

A civil society response needs to enable risk taking when developing detection technology. Within any particular deployment of the system, analysts and technologists will be under both time and financial pressure. They will have to make decisions regarding the technologies that are used, and how they are implemented, which will tend to prioritise safer, tested and better understood approaches over riskier and newer alternatives.

## About the Institute for Strategic Dialogue

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