DIRECTING RESPONSES AGAINST ILLICIT INFLUENCE OPERATIONS (D-RAIL)

by Carl Miller

Directing Responses Against Illicit Influence Operations (D-RAIL)

Part 1. The problem

The fact that disinformation poses an existential threat to democracies is one of the most important of our age. Civic societal coalitions have sprung up to fight it. Summits are convened. New government task forces have been created, as have academic centres, regulatory groups, and a fledgling industry of private-sector start-ups; all to make sure that disinformation doesn't win.

All of which makes this article a difficult one to write. Defining the information threat that democracies face as primarily one of disinformation is, as I see it, sending us off in the wrong direction. Rather than 'disinformation', the information threat that democracies need to counter is something different: orchestrated, professional, sometimes covert influence campaigns mounted, chiefly, by their geopolitical adversaries. It is these campaigns that deliberately corrode democratic institutions and processes. And whilst these campaigns sometimes do spread disinformation, that is just one tactic amongst many. We need to step away from trying to have a general response to lies circulating around on the Internet, and towards a strategy to specifically disrupt the money flows, people and technical infrastructure that facilitate these campaigns. That strategy is what this article will suggest.

First, however, we need to address this question of terminology. There has been an enduring, hugely influential and mistaken belief that illicit forms of influence online are somehow synonymous with untruths. This gained momentum during the 2016 U.S. Presidential cycle when the phrase 'fake news' began to be used to describe, at first, spoofed news spam sites. The definition widened, though, to encompass an ever-broadening array of things that were often neither fake nor news. The platforms themselves have generally preferred terms like 'coordinated inauthentic behaviour' to focus on behaviour rather than content, but this has tended to remain a *lingua franca* mostly popular within platforms and the expert community. For the rest of us, the word that has really stuck is disinformation.

Focusing on the content veracity is actually a poor shorthand for describing what we need to care about. The vast majority of actual online false content – that is, someone lying to someone else – isn't something worth spending already limited resources to respond to. Someone deceiving users about how tall they are on Tinder, or how rich they are on Quora might be reprehensible, but surely is not something we need a grand all-of-society response to.

Much of what does threaten democracies also aren't lies. As we started pulling illicit campaigns apart, myself and my colleagues began to realise they used a whole variety of tactics which had little to do with truth or lies at all. Campaigns were far more likely to affirm the existing beliefs of their audience rather than try to change them with disinformation.

Some build a distorted picture of the world simply by amplifying some truths over others. They often sought to gain influence by building different senses of patriotism, or masculinity and new senses of identity in the audiences they could reach. Extolling the strength of Putin could be done through bear memes. Attacking the authority of experts was done through cherry-picking errors in scientific journals and making sure target audience sees little else. Hostile states will covertly amplify conspiracy theorists, while suppressing voters, activists and journalists with threats of violence to control the information environment.

The European External Action Service has pointed in the right direction with their coinage of the acronym 'FIMI' (Foreign Information Manipulation and Interference).The creation of the term FIMI has lead to an increased focus on manipulation and interference, shifting the focus from illicit behaviours rather than untruthful content and, even further. the tangible campaigns – which we can call 'illicit influence operations (IIOs) – that underlie the threat actors' behaviour. A clunky term, I know, but what the phrase points to is the existence of an *operation*; a series of activities coordinated over time in a conscious way. They are systematic and deliberate, conducted by specialists – often professionals – according to techniques whose roots can be traced back to a number of historical antecedents, from the doctrines of Active Measures and Reflexive Control of the Cold War to the political and psychological warfare of the First and Second World Wars.

For illicit influence campaigns, information itself is purely an instrument; a way of causing some sort of changed behaviour in an intended audience. If propagating a lie is the way to achieve that, these campaigns will do it. But they will also amplify truths, confirm the reader's existing beliefs, try to build friendships, trigger outrage, create financial incentives and change social norms to do so, too. There are a great many ways to influence someone without lying to them. To call these campaigns disinformation is to confuse a single tactic with the overall strategy. It is the same as confusing a bayonet charge for all of warfare.

The reason I labour this point is that defining any problem matters enormously in determining the solutions you then reach for. Dealing with disinformation naturally points towards a series of responses to do with identifying and correcting falsehoods: digital literacy training to help people spot lies, supporting journalists and fact-checkers to call out the lies, and increasing pressure on platforms to take the lies down are all essential measures. These responses are surely important for societal health in general, but they will, by themselves, not protect democracy from the sorts of illicit campaigns we've just discussed. Calling out disinformation alone will not deal with all the other ways this tradecraft works.

Yet defining the problem as influence *operations* leads us to focus on the actual people, infrastructure, organisation and money that make up the operation. And this is the way forwards: to bring far greater focus towards disrupting illicit influence campaigns themselves. We need to focus on how to make these operations harder to do, less profitable, less successful, more difficult, chaotic and more prone to error. We will never stop bad actors conducting IIOs. But like

anyone else, they do not have infinite resources. They, too, fight for funding against competing priorities. The future ability of our societies to endure these sorts of operations will depend on our capacity to drive up the cost and difficulty to do illicit influence.

The question of *how* brings us to the second part of this article, and that is to propose a process to undermine IIOs. It is based on a simple premise: influence is harder to have than to stop. Successfully influencing a particular audience in a dependable and consistent way is actually very difficult and expensive. To do so, the adversary needs to buy assets, create identities, coordinate, plan, learn, and ultimately capture and exploit the attention of targeted audiences in ways that can reliably conduce the behaviour they want. All of that can be disrupted.

The process detailed below is therefore a strategy to cause IIOs to make poorer decisions, become riskier, more expensive, more prone to error, harder to coordinate and more burdensome to sustain. It aims to make the audiences they target harder to reach, less likely to pay attention, and less likely to behave in the way the adversary intends them to act. If we can find ways to do this, we prevent illicit forms of influence having the effects they want, for a fraction of the cost it takes to conduct them. It is called Directing Responses Against Illicit Influence Operations, or D-RAIL.

Part 2. A solution

The point of D-RAIL is to introduce more friction, error and failure to the campaigns that are promulgated to undermine democratic institutions. As far as possible, it is data-led; it creates room for creativity and new ideas; it does not focus on any particular platform or form of influence, and it is designed to constantly change and to progressively improve over time.

Now, I know what you might be thinking: 'not another framework!'. The defender community is, I know, making efforts to harmonise the methodologies, taxonomies and definitions that we have, recognising that it's actually helpful to have fewer of these things rather than more of them. I'd urge you not to see D-RAIL as a framework – rather, it's a strategy for connecting and using the frameworks[1] that we already have. It connects three, key frameworks together:

1. To describe the illicit influence operation (IIO): the Cyber-operations Kill Chain. Ben Nimmo and Eric Hutchins proposed an 'Online Influence Operation Kill Chain'. An idea itself borrowed from cyber security, their kill chain suggests that there is a meaningful series of phases that a influence operation necessarily has to go through. This is an extremely useful way of thinking about influence operations, and I relied on a slightly more generalised version of Nimmo and Hutchins' kill chain, below.

2. To source and describe ways of disrupting an IIO: the DISARM Blue Framework. DISARM is an open framework to develop a shared and systematic way of describing disinformation incidents and responses. The Blue framework collects together a very large number of responses which are described in a standardised way as 'tactics, techniques and procedures' (TTPs). Caution is advised when using the framework, as an "alternative based on democratic values and ethical principles" is being developed.

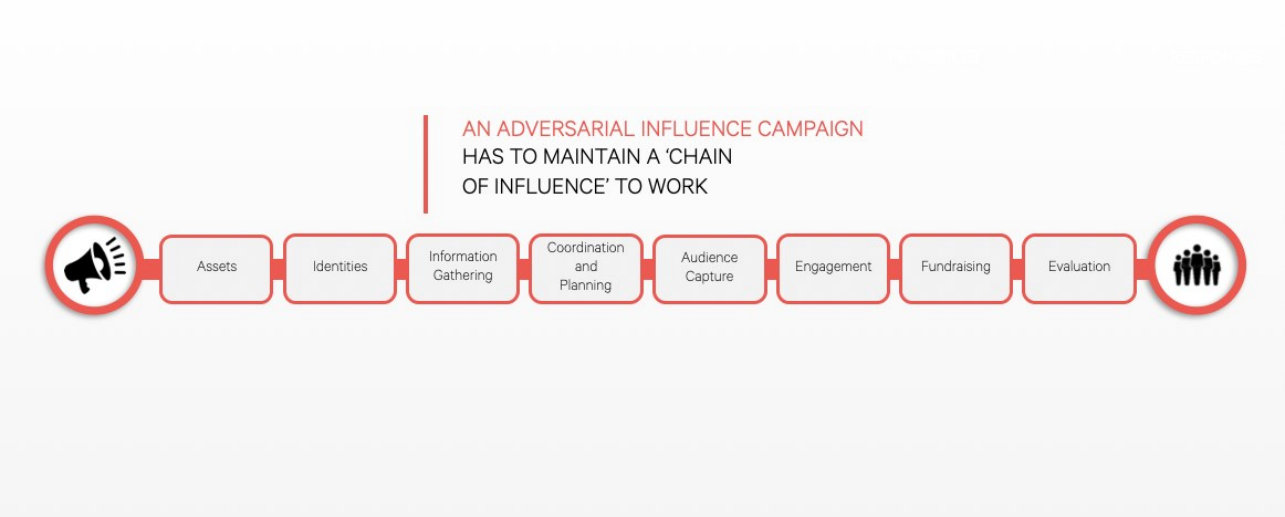
3. To evaluate responses: EU DisinfoLab's emerging cost-effectiveness evaluation framework. The entire point of D-RAIL is to develop responses that inflict greater costs on illicit influence campaigns than are caused by the responses themselves. EU DisinfoLab has developed a cost-effectiveness assessment method that suggests considering the expertise, time and financial resources invested by the defender community in relation to the goals achieved by the responses given to a campaign. If the ultimate goal (the disappearance of the campaign from the public space) is not reached, the effectiveness is measured by evaluating five intermediate objectives: the increase in situational awareness (relying on stakeholders gathering, sharing and amplifying knowledge on the operation); the impact on the threat actors' capabilities in content production and infrastructure distribution; the triggering of new responses by the defender community but also by the threat actors; the increase in attribution opportunities (which

[1] I.e., structured sets of guidelines designed to systematically address specific problems efficiently and consistently

could trigger other actions, such as sanctions), and the deterrence effect on the threat actors from continuing the campaign (for instance, by disappearing from a given country or platform).

With these frameworks in mind, D-RAIL consists of five steps. The overall idea is to describe the IIO as a chain of influence, contour responses against it to break that chain of influence, and then constantly measure and evaluate both the malign operation and your attempts to disrupt it over time, and learn.

Step 1. Articulate the illicit influence operation's 'chain of influence'



Every illicit influence campaign has an actor that conducts it, an audience it targets, an information environment with which it reaches that audience, an effect it intends and, presumably, an underlying aim or interest in mind. To do any of this, campaigns must have and do a number of tangible things in a particular order, which can be articulated as a chain of influence. Some links of this chain will be more visible and others less, but the idea of a chain is that there are a set of meaningful commonalities in phasing and deployment that illicit campaigns need to go through, regardless of the forms of influence they seek to use.[2]

Link 1. Acquire assets. Any operation has to acquire assets in some form. This may include IP addresses and email accounts, social media accounts, TV stations, physical offices, physical servers, crypto-currency wallets, VPN accounts, the registration of websites, camera equipment, registered companies, political parties and so on. These can be openly purchased, or acquired through corruption or coercion.

[2] This is a modified version of the Online Influence Operations Kill Chain proposed by Ben Nimmo and Eric Hutchins, adaptedto be made more general for other kinds of influence operation too. Moreover, these are combined with the various phases identified in the DISARM framework to emphasise interoperability.

Link 2. Establish identities. The campaign next needs to acquire an identity(s) that the operation will use. This can be overt, through the creation of company branding, or the building of a political or theological reputation of an identifiable group. It can also be covert, through the use of stolen pictures, the creation of front companies or by compromising websites or social media accounts.

Link 3. Gather information. The campaign needs to learn about the audiences they want to influence. This might include research regarding their attitudes, the information environments they inhabit, and the sorts of language they use.

Link 4. Coordinate. Influence activities will necessarily involve resourcing, management, and some form of internal or external coordination. This coordination apparatus may take many forms, from instructions posted in Telegram channels, targets on Google spreadsheets to military chains of command.

Link 5. Capture target audience. One of the most formidable challenges faced by any influence campaign is to win and hold the attention of a busy, often distracted target audience. The adversary may have to develop fanbases, channels and groups through advertising, recruitment, messaging, the manipulation of search engines, door-to-door promotions and so on.

Link 6. Exploit captured attention. Only once the previous steps have been completed will the adversary be able to exploit the attention of the targeted audience. This will inevitably involve some kind of theory of influence – an idea that changing the target audience's information environment in a given way will producea desired set of resultant behaviours. This may involve the propagation of disinformation, but the operation might target really any part of the intellectual, moral and psychological worlds of the target audience, from what they think is true to whom they trust, the emotions they feel, the identities they hold, the friendships they cherish, the sense of grievance and injustice they feel and many, many others.

Link 7. Ensure longevity. Operations may conduct activities that ensures the continuance of the campaign itself. This might take the form of fundraising, whether through reusing technical assets for cybercrime, monetising the attention of the target audience, selling products, or seeking money through economic corruption or coercion. Other activities might work to ensure technical longevity (such as maintaining network access), or the maintenance of identities through, say, dissuading an investigation that may expose them.

Link 8. Learn and evolve. Finally, the adversary will likely try to evaluate the effects of its own campaign in order to justify its budgets and risks, and to evolve the campaign in the face of new circumstances, failures, pressures, social

trends, technological adoptions and so on. The DISARM campaign lists a number of specific activities here, from social media engagement to the measurement of performance.

Step 2. Create counter-activities to break the illicit chain of influence

The second step is to elaborate a series of counter-activities to introduce as much as friction, error, risk and failure into the IIO as possible. It is inevitable that a combination of data, judgments, assessments and estimates will come together in any attempt to describe any chain of influence. Once this is done, a 'critical vulnerability analysis' can be conducted to try to identify the parts of the adversary's chain of influence that could be disrupted. This analysis might include the following considerations partially drawn from the APEASE model:

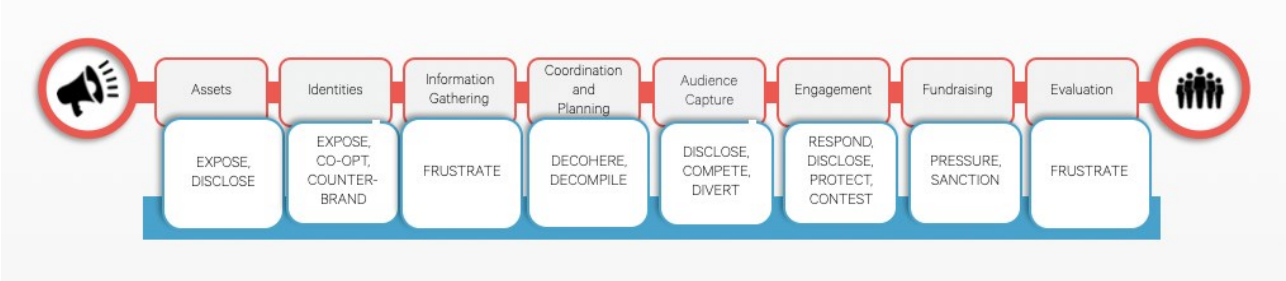
- Acceptability. Whether the counter-response accords with the ethics and values of the democracies they are meant to protect (see the section on ethics at the end of this essay).

- Practicability. Whether an intervention is likely to be delivered as planned.

- Effectiveness. Whether an intervention is likely to achieve the outcomes at the scales and speeds needed.

- Affordability. Whether it can be implemented within the available budget.

- Spillover effects. Whether the intervention is likely to have unintended effects.



Of the great number of possible counter-activities that may be conducted, some are suggested below although please note, this is not to particularly endorse any. Every response should of course be thoroughly subject to both risk assessment and the legal and ethical frameworks of whatever organisation is conducting them.

Breaking Link 1, when the adversary acquires assets. One may expose any that are hidden, inauthentic or acquired via illicit means. Influence assets – when being used abusively – can be systematically reported to the relevant intermediary/intermediaries, who can take action if required to by law or, if they wish to do so, on the basis of their terms of service. Typically relevant intermediaries would be hosting service providers, online platforms such as social media and domain name registries and registrars.[3]

Breaking Link 2, when the adversary creates identities. In general, this is an opportunity to expose the use of illicit or covert identities that may be used for illicit influence (and act).Responses can entail public revelations of the true intentions underlying deceptive identities, or more targeted communications, such as informing co-opted voices (such as influencers, researchers or journalists) of the real interests behind a given campaign. Moreover, impersonation and identity theft (including digital identity) can constitute criminally prosecutable actions in some jurisdictions.

Breaking Link 3, when the adversary gathers information. This is a difficult link to disrupt, but may include denial of the adversary's access to research and polling organisations (via 'know your customer' obligations). It may also include the promotion of privacy-protecting behaviours and practices amongst audiences that are targeted by IIOs.

Breaking Link 4, when the adversary coordinates. Some influence campaigns have significant coordination hurdles to surmount, and this provides an opportunity to increase the coordination costs between these actors. Influence campaigns might try to galvanise different groups through forums, open channels, and shared documents, but such coordination efforts are difficult to detect in the early stages. However, once the operation is underway, it is possible to expose Coordinated Inauthentic Behaviour (CIB), involving platform providers, and even regulators or law enforcement agencies. Influence campaigns might establish commercial relationships with illicit suppliers, such as troll-farms or influencers, and these can be targeted either by the platforms under their own definition of CIB or possibly with legal action where the activity itself (such as in the case of harassment) is illegal. Responses introducing coordination costs require, coordination and cross-platform information sharing will be vital to identify when illicit influence campaigns are coordinating.

Breaking Link 5, when the adversary captures and holds attention. One of the most tenuous and difficult stages in any influence operation is often capturing the attention of target audiences. It is likely during this phase that the influence operation will be its most visible and, possibly, vulnerable. As illicit campaign can and often do use advertising, it is essential for advertising platforms and payment service providers to have strategies in place to prevent

[3] Including, now, via the DSA Illegal Online Content Reports.

their abuse, such as short delays before publication that allow verification before harm is done, and comprehensive and rigorously enforced rules against abuses of this kind.

Breaking Link 6, when the adversary exploits attention. This is the part of the operation where the campaign actually engages in information manipulation and where their activities are therefore necessarily most visible. Counter-responses range from the extremely tactical and responsive to years-long undertakings to contest ideas such as patriotism or duty.

- The short-term and most reactive counter-responses is counter-messaging. This might involve (echoing the language of DISARM) the reduction of 'polarisation by connecting and presenting sympathetic renditions of opposite views" [C00111], "Engage payload and debunk" [C00119], or "Develop a compelling narrative (truth based)" [C00030]. As DISARM notes, counter-messaging requires extreme prudence to avoid misleading interpretations and the operators' potential instrumentalisation of these messages." The protection of activists, journalists and politicians targeted by illicit influence campaigns by harassment and threats is also a time-sensitive response that should be done.

- In the medium term, one can also more proactively pre-bunk illicit messaging, and also expose the fictitious sources of authority and legitimacy that the adversary might create.

- In the long term, there are also responses that are less about directly responding to the specific messages they send, and more about taking the initiative to contest and replace the more general framings and narratives, identities and lexicons that the adversary promotes.

Breaking Link 7, when the adversary ensures longevity. Undermining the activities that ensure the continuance of illicit influence campaigns are, by definition, amongst the most effective responses that we can make. Different stakeholders can take action. Platforms have different approaches to demonetising content that fails to comply with their guidelines. Detecting and addressing these financial challenges is a key element of investigating these campaigns, despite threat actors' attempts to conceal their infrastructure to continue operating. If advertisers monetise their audiences through programmatic advertising, their assets can be reported to indices that prevent ad sales to disinformation sources. Organisations such as the Global Disinformation Index and NewsGuard seek to hinder financial incentives to disinformation by discouraging advertisers from supporting deceptive sources. One might also undertake consumer-facing activities to reduce the adversary's revenue generation, through boycotts and campaigns.

Breaking Link 8, when the adversary learns and evolves. In many cases, the adversary will try to learn how well they are doing. This will by nature likely be less visible, but any engagement of the adversary with reputable research or polling organisations, for instance, should be disclosed to those organisations. Understanding the adversary's TTPs

and their evolution is fundamental. As seen in the Doppelganger, operation, malign operators adapt by improving their obfuscation techniques – finding ways to circumvent deplatforming and sanctions – leading to a higher proliferation of disinformation.

Step 3. Continuously monitor and evaluate both the illicit campaign and the responses to it

The *raison d'être* of any illicit influence campaign is to achieve behavioural and attitudinal effects in a target audience. Likewise, the whole point of any response to any illicit influence campaign is to prevent these effects from occurring and/or to increase the costs whereby they are achieved. And so, the point of D-RAIL is to create a system where the possible effects and harms caused by the adversary can be judged, as can the attempts to stop them. Any process to disrupt illicit influence has to be ruthlessly empirical. The illicit influence campaigns will, of course, change and evolve, and as it does so, the counter-responses must subsequently evolve. D-RAIL must therefore continue to make a series of continuous measures through a combination of data-driven measures, focussed investigation, judgements and assessments.

Methodologies such as EU DisinfoLab's are being constructed to measure the manifold impacts of IIOs and the cost-effectiveness of attempts to disrupt them, and I do not intend to repeat them here. As these methods become more generalised and widespread, they should be deployed to answer three key questions at the heart of D-RAIL:

What are the activities likely being conducted by the illicit influence campaign? This is a continuous monitoring of anything that can be known about the adversary's activities that are done to construct, sustain or evolve its chain of influence. This will range from identifying assets they might have registered, to analyses of advertising activity and fundraising. This form of monitoring will likely cover very different scales and natures of data and methodologies to analyse them, from big data analysis to focused OSINT investigation. For this very reason, data harmonisation is going to be essential, and defenders across industries and fields must begin to migrate to common language, coding systems and repositories such as OpenCTI.

What is the impact of the illicit influence campaign? These are measurements of the effects of the illicit campaign on their audience. These measures are likely to be the most uncertain and also important, and vary from observed behaviours to the polling of attitudes. It is always difficult to know what causes a particular behaviour or changes an attitude, especially in a world full of competing influences. However, they are important because they define the ultimate harm that the adversary has inflicted.

Are the responses succeeding? The last series of measures will relate to the possible effects and outcomes of D-RAIL. These might be include the costs imposed on, and assets removed from, the IIO. It also may cover efforts to expose the activities of the illicit influence operation, such as the proportion of their intended audience that have become aware of a covert identity it is using.

Step 4. Learn and Evolve

Finally, D-RAIL has to improve over time faster than the IIOs it confronts. To become iteratively more effective over time, key points of learning must be translated from one to the next. The final element of D-RAIL is therefore to create more general, strategic understandings of illicit influence campaigns and the attempts to stop them through the creation of two bodies of evidence.

1. A database of IIOs. This must be a standardised a list of detected illicit activities, across the entire chain of influence: known technical assets, actors, companies, financial assets, technology and so on. It will also include known illicit tradecraft: their TTPs, operational practices, methods of obfuscation and any other behaviours that are associated with their chain of influence. For example, the defender community is working on initiatives such as the FIMI Information Sharing and Analysis Centre (ISAC) to collect, analyse, and share knowledge of IIOs in a standardised manner.

2. A repository of evaluated responses. The evaluation of counter-influence responses should be captured in a second database. This will build a long-term institutional memory of which responses have seen success or failure, the contexts associated with that success or failure, and the costs and risks associated with the response. This will iteratively increase the evidence-base that can be drawn on when deciding which counter-reposes to deploy in any given context.

Part 3. Capability and ethics

The suggestion of any kind of a method like D-RAIL inevitably begs the questions of how the defender community needs to be structured and equipped in order to be able to actually do it. That is a good question, and one both I and others have tangled with for many years, and whilst I won't repeat any of those efforts here, it is worth pulling out a few key thoughts on how to make D-RAIL a practical reality.

Diversity. The reality is that to be effective, D-RAIL will require capabilities – the permissions, powers, codes and technologies – that are far from being found in a single organisation, and that are, in fact, scattered across society. Journalists, researchers, and OSINT practitioners will be able to conduct investigations no-one else can, especially into attribution (see below). Some responses will only legally be possible for governments, whilst civil society will be able to reach vulnerable and targeted communities in ways that governments simply can not. Coalitions will therefore be necessary across all of these different groups to create portfolios of responses that encompass many different strengths and skills. Despite effectiveness of responses within the defender community varying broadly, community engagement stimulates situational awareness, which can trigger a positive spillover in stakeholders' actions.

Standardisation. Precisely because responses must be diverse, they also need to become much more standardised. This is why, as far as possible, I have sought to have D-RAIL reflect the languages, standards, frameworks and definitions that are emerging as consensual. Harmonised TTPs framework such as DISARM, threat intelligence operationalisation such as STIX language, and platforms that could serve as repositories such as OpenCTI are vital to the way forward.. Common languages and codes around impact, effectiveness, cost, and risk will be essential to join the diverse range of organisations involved.

Attribution. D-RAIL is predicated on the idea that it is possible to actually reveal the underlying attributes of the illicit influence campaign: who they are, what they are doing, and whether it is working. This is extremely difficult to do simply through the analysis of online content itself, and in terms of investigatory emphasis, D-RAIL is an argument for greater focus and information-sharing on attribution. This includes the more intense involvement of journalists, researchers, and OSINT practitioners seeking to uncover campaigns, and also more information shared by platforms that can aid investigators in making attribution claims (albeit rarely conclusive ones).

Management of uncertainty. For the reasons given above, D-RAIL will also need to cope with uncertainty in the information it gathers and uses. Uncertainty must be managed and measured, and cannot be either ignored or avoided.

Emerging technology. Alongside uncertainty, however, information defenders will also have to cope with very large amounts of data that is dynamic, possibly contradictory and very rich. They will also have to contend with tradecrafts of

illicit influence that will experiment with and try to use emerging technologies wherever they can. Effective responses will therefore be required to be driven by the constant use of emerging technology for detection and assessment.

Data Access and Sharing: Nothing is more essential to effective responses to IIOs than the ability to acquire sufficient information about them. In part, this will depend on information sharing from online service and platform providers, either voluntarily or via the emergent regulatory regimes, such as, inter alia, the advertising transparency and data access and scrutiny requirements of the EU's Digital Services Act.

Ethics. Anyone who has read this essay will, I'm sure, feel that there are many questions left unanswered. Who is intended to run D-RAIL? What ethical framework is supposed to govern it? Am I calling for us to use the enemy's tactics against them? The exposition of a full and detailed ethical framework would require at least the length of this essay again. However, there are a number of key principles that should shape the activities to disrupt any illicit campaign.

- Minimise collateral intrusion. One of the main aims of D-RAIL is to pivot responses away from broad, population-level effects, and towards those specifically targeting illicit influence campaigns. This is in order to minimise the number of normal people who are affected in any way, either by the illicit influence campaign or the attempts to stop it. Targeted responses should be preferred over more general ones.

- Necessity and proportionality. The justification for any use of D-RAIL is the harm that is being caused by the illicit influence campaign. The harms should therefore be necessarily severe in order to warrant a response, and the responses made, in terms of cost and risk, should always be proportionate to the harm they are attempting to prevent.

- The inclusion of non-informational responses. Much of what I emphasise in this process is the importance of fighting information operations *asymmetrically*, outside of information spaces. The point I am making is that tangible campaigns sit behind much of the most damaging and visible information we encounter. These campaigns have people, assets, financial infrastructures, budgets and brands that can be targeted by activism and the law, by sanctions, take-down, de-listing, and exposure. When we focus on the operations that manipulate information spaces – the behaviour rather than just the content – a whole range of responses become apparent that do not require us to weaponise information spaces ourselves.

- An avoidance of inauthentic, harmful, fictitious or untrue responses. There is sometimes the suggestion of using some of the techniques of information operations against information operations. However, this is likely to be a losing strategy for democracies. Liberal democracies have most to benefit from information spaces that are protected, and the most to lose from information spaces that, when weaponised, are neither used nor trusted. Fighting deception with deception is a poor strategic choice for liberal democracies overall.

- Responses must respect the wider web of democratic values they seek to protect. This includes consumer protections, the freedoms of speech and assembly, and the important frameworks that already exist around data protection and integrity. Here, I argue, by mounting much narrower responses against specific illicit influence campaigns rather than disinformation in general, responses can often avoid becoming entangled in much wider questions to do with, say, the balancing of the freedom of speech with the removal of online harms.

I'll end with a final thought. To respond to IIOs, we need to blend invention and imagination with a ruthless and honest empiricism. We need a profusion of new ideas for how to cut chains of shadowy, unaccountable and covert influence, and also ways of telling which actually work. It is only then that will be able to become more innovative than those who run these sorts of operations and that defence will draw ahead of offence in this strange conflict over, about, through and within information itself.