DISARM Design Guide

Contents

1 Introduction

What this report covers, and who's responsible for that.

1.1 About this Report

This report covers the motivations, philosophy, use and intentions for the DISARM disinformation models. It's one of a series of reports on how information security principles and practices can be used to improve our understanding of cognitive security and improve responses to information operations, and specifically to disinformation campaigns and incidents.

The original brief for the report was to create a 1-2 page description of each tactic and technique in DISARM (the 'squares' on the DISARM grid) for practical use. This was to include advice on how to investigate each tactic - e.g. "This incident is using the pump priming tactic, what are the indicators you need to look for, what are the countermeasures you could use”", and how to counter it in the user's context.

The DISARM models are open source, licensed under CC-by-4.0 (Creative Commons Attribution-ShareAlike 4.0 International).

1.2 Structure of this report

This report is structured around sections on DISARM design and use, and can be summarised as:

- This introduction section: history of DISARM's creation
- DISARM design and philosophy: why and how we built the frameworks, and the design choices we made
- DISARM component designs

- Using DISARM: Tools, technique and suggested uses
- Future work: notes and ideas for improving this work

This is a companion document to the DISARM TTP Guide and the master copy of the DISARM models, contained in github repository https://github.com/DISARMFoundation.

This report is a living document. This is version 1.2 based on the renaming of AMITT to DISARM.

1.3 Disinformation Defence

State actors, private influence operators, and grassroots groups are exploiting the openness and reach of the Internet to manipulate populations at a distance. This is an extension of a decades-long struggle for “hearts and minds” via propaganda, influence operations, and information warfare. Recent advances include computational propaganda: the use of algorithms, including machine learning and artificial intelligence, in online manipulation.

There are many definitions of misinformation, disinformation, incident etc. and teams dedicated to improving them. This report uses these working definitions:

- Disinformation is the deliberate promotion of false, misleading or misattributed information, usually designed to change the beliefs or actions of large numbers of people, as a tool to help meet an exterior goal.
- Misinformation is false or misleading information that’s potentially harmful.

The structure and propagation patterns of misinformation attacks have many similarities to those seen in information security and computer hacking. Analyzing and building on similarities with information security frameworks gives defenders better ways to describe, identify and counter misinformation-based attacks.

1.4 DISARM

DISARM is a set of data standards and an open source knowledge base of both red team and blue team disinformation tactics and techniques. DISARM’s intended users are disinformation responders; its purpose is to give them the ability to tactically respond to disinformation incidents, to plan defenses and countermoves, and to transfer information security principles to the disinformation sphere. It provides a common taxonomy for cognitive security offense and defence, a framework to rapidly share threat intelligence, and a conceptual tool for strengthening disinformation defences through red teaming, risk analysis, replays and simulations.

DISARM consists of blue team (defence) and red team (attack) models, and a repository of descriptions, mitigations and examples. To create DISARM, we placed misinformation components into a framework based on standards (including ATT&CK and STIX) commonly used to describe information security incidents. DISARM frameworks are designed to fit the same toolsets and use cases that STIX and ATT&CK are designed for.

1.5 Acknowledgements

DISARM was originally developed in 2019 as AMITT by the Credibility Coalition's Misinfosec Working Group (MisinfosecWG), with inputs from the misinfosec community including experts who generously gave up a day of their time to workshop the first framework (the Atlanta workshop in 2019), then 2 days of their time to workshop potential counters (the DC workshop in November 2019).

The MisinfosecWG brainstormed, collated and devised new ways to counter or mitigate online manipulation, focusing on manipulation through disinformation and its known and potential countermeasures and mitigations. Our intent was always to give responders the ability to transfer other information security principles to the misinformation sphere, and to plan defenses and countermoves. For instance, we started the disinformation countermeasures workshop in Washington DC, with two main goals:

- Create the first version of a disinformation "Blue Team" playbook. For defenders, information security people and organizations, this will be a set of responses to misinformation attacks—the networks, the response types, the frameworks, and examples.
- Define how to support an operational global MisinfoSec_ISAO network. For potential response center participants and leaders, this will be the process, methods and understanding needed to connect, including suggesting partners, collaborators and funders.

MisinfosecWG was a short-term project to create information security-inspired standards for sharing information about misinformation incidents and how to respond to them. It was succeeded by the CogSecCollab nonprofit, which maintained the AMITT standards, and then the DISARM Foundation, which currently maintains the DISARM standards, with Sara-Jayne Terp and Pablo Breuer acting as design authorities.

CogSecCollab extended the original AMITT work, adding disinformation tools to infosec incident response and information sharing tools including MISP and STIX, and trialling the use of AMITT in its prototype disinformation Security Operations Centers (SOCs), including the CTI League's disinformation team, and with organisations including NATO, the EU and disinformation teams from several countries. Under the tutelage of the DISARM Foundation AMITT was renamed to DISARM in 2021.

CogSecCollab was in discussions with MITRE about the MITRE team taking on AMITT alongside the ATT&CK model which inspired AMITT's design. MITRE created a fork of AMITT called SP!CE with the help of Florida International University. Following discussions between MITRE and the DISARM Foundation the two organizations collaborated to merge SP!CE and AMITT to create the DISARM set of frameworks.

It takes a village, and we have many people to thank for their contributions to DISARM. Thank you all. We hope, with the new work, that we've done you proud.

2 DISARM Toolset Design and Philosophy

The DISARM toolset was created from a need for a common language for disinformation - at its creation time, our community included media, academics, infosec professionals, data scientists, government and people from other disciplines who all had different words for disinformation concepts and objects. DISARM tools should ideally provide a way for people from different fields to talk about misinformation incidents without confusion.

This section covers why and how we built the DISARM toolset, how its models connect to each other, and the design choices we made in their creation.

2.1 Disinformation as an Ecosystem

Our first move, back in 2016, was to talk about disinformation not as an isolated “fake news” problem, but as an ecosystem in which multiple actors with different motives (geopolitics, power, money, attention) interacted with misinformation and information flows, stories, beliefs, communities and individuals, websites, media, platforms and algorithms.

That was a lot of moving parts, and a lot of data, so we looked at other entities that were analysing ways that online and community beliefs and emotions could be changed, or analysing attacks on flows of information across the internet. These existing communities included social science, online marketing, adtech (online advertising technology) crisis data mapping, information security and data science. Our team all came from different directions on this, and all had different words and models for the same concepts, so in 2018, we formed two groups to connect them, and create a common language to talk about disinformation.

2.2 Connecting Defence Actors

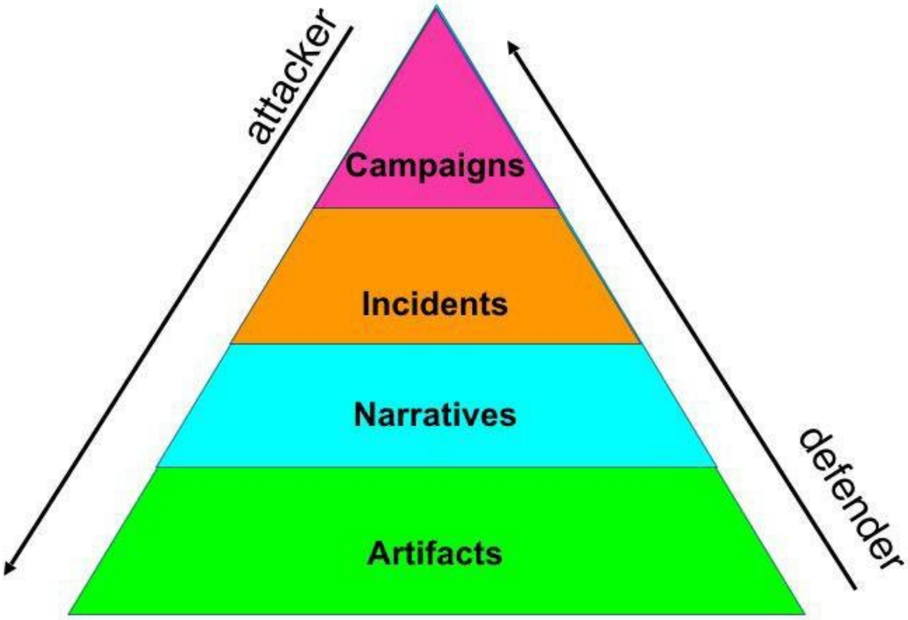
When we started, we knew that our best chance of creating good disinformation defences meant connecting together people from very different worlds:

- The information operations specialists who spent their days analysing “conflict short of war” - military techniques like psychological operations (“psyops”), and other power moves between nationstates
- Data scientists, who analysed sets of objects and flows of information across the internet using techniques like machine learning and social network analysis to pick apart patterns of accounts, text, hashtags, urls, groups, and the relationships between them all.
- Social scientists and psychologists who studied human cognitive vulnerabilities,

group dynamics, and the flow and effect of narratives on beliefs and emotions.

- Information security (infosec) experts, who had already built tools, techniques and processes to protect information held in very similar topologies, which instead of being communities of people were networks of machines.

Our first model, the disinformation pyramid, was built to help these groups talk to each other.



Disinformation Pyramid

Here we're looking at the different views that creators of misinformation ('attackers') and the people trying to counter them ('defenders') have (a third group involved, the targets of the misinformation, 'populations', aren't part of this diagram).

- Disinformation creators often persist in the ecosystem, focusing on one or more longer-term objectives (e.g. destabilize French politics, or reduce vaccination rates in target countries). We called these longer-term objectives "**campaigns**"; Clint Watts
labelled these longer-term actors "advanced persistent manipulators" (APMs), mirroring the infosec term "advanced persistent threat". Many APMs are nation-state actors, using disinformation to attack other nations: this is the pyramid level that many of the information operations specialists were working at.
- **Incidents** are shorter-term sets of disinformation activity, often around a specific topic or event (e.g. Macrongate). These bursts of activity might be triggered by an event or

opportunity to make money (there are many opportunists pushing misinformation), or they might be the result of a team of people working towards a desired effect: a change in beliefs or emotions relative to a specific person, group, object, concept, or event; or a weakening of an opposing group, belief etc by creating chaos and confusion. Campaigns typically contain multiple incidents, sometimes happening at the same time. Information security experts recognised this level of attack and mitigation as similar to the work they did deterring and mitigating attacks on information systems.

- **Narratives** are the stories that we base our beliefs on: "identity narratives" about who we are, "in-group" and "out-group" narratives about the groups that we do and don't belong to, and other narratives about what's happening in the world around us. Most incidents use and rest on narratives, and we found ourselves tracking and talking about these as a useful abstraction of all the artifacts we collected for each incident. Narratives scientists fit into this layer of the pyramid, and it was a useful level to bring in social scientists and psychologists.
- **Artifacts** are the messages, images, accounts, relationships, and groups that a disinformation actor uses to create narratives and incidents. Artifacts are visible in each incident, often in large volumes, and are the disinformation layer that data scientists and other data specialists usually worked on.

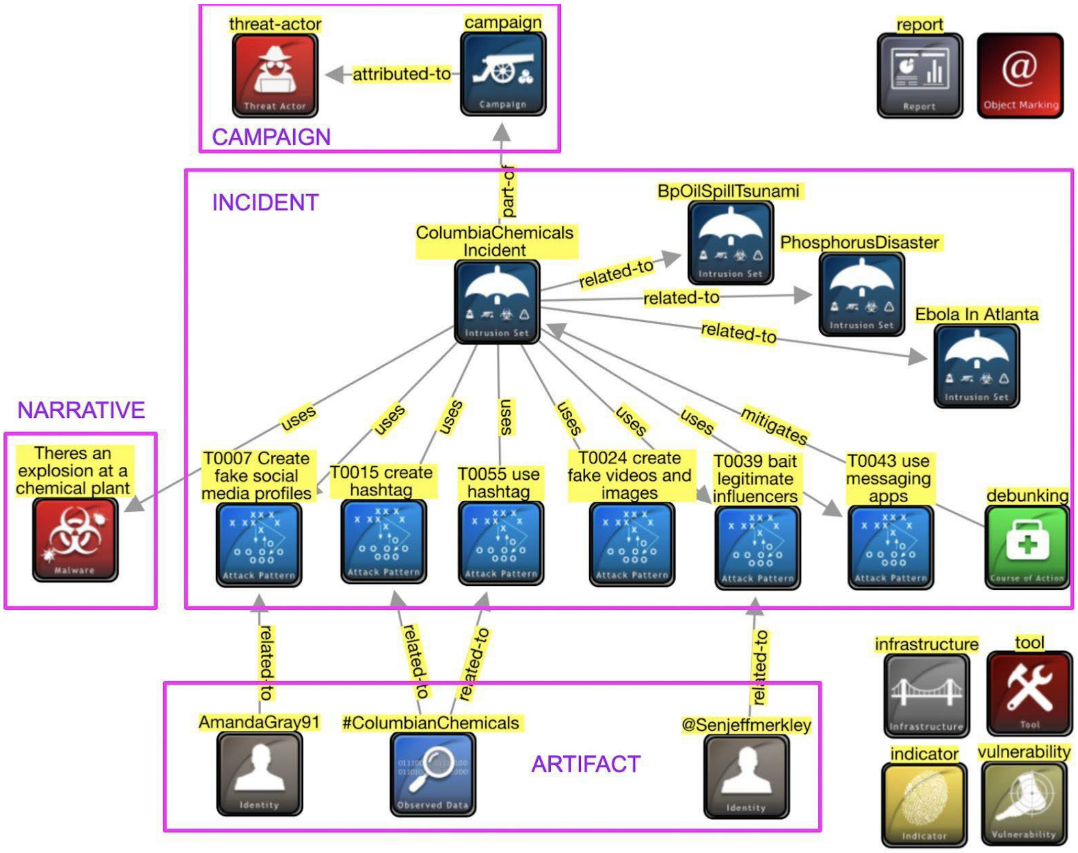
While the attacker sees the whole of the pyramid from the top down, the defender usually sees it from the bottom up, working back from artifacts to understand incidents and campaigns, unless they're lucky enough to have good insider information or intelligence, or have kept databases of information in forms that can be used to anticipate and compare artifacts, narratives etc to earlier work.

When we drew this pyramid in 2018, most of the misinformation tracking work that we saw was at the artifact level, with some work on the narrative (story) level, with Pablo coining the phrase that we were "admiring the problem" and needed to move to defence. Today (2020), all levels are investigated and connected, and the wider conversation has moved from tracking to defence and mitigation. We've also started to see human-readable reports on disinformation events that follow the layer model structure - starting with the wider context including campaigns, then an incident description including techniques used, then a list of narratives, and artifacts of interest at the end.

2.3 Component-based disinformation models

A useful view of a disinformation incident is as a collection of the objects seen within it, and the relationships between them. Many disinformation researchers already organise their information this way (as do the OSINT, intelligence and journalism-inspired research that much of this work is based on), with some of our earlier collaborators going as far as building “murder walls” to track groups and incidents. These are formalised as sociotechnical systems models - models of the complex networks of interacting communities, accounts and technologies that make up a disinformation incident or campaign.

The infosec community already has a data standard for this, STIX https://oasis-open.github.io/cti-documentation/, which also comes with a standard, TAXII, for how to share STIX data across systems. MisinfosecWG created a disinformation version of STIX, mapping its existing object types for disinformation use, and adding two new STIX object types: incident and narrative, because the existing objects, intrusion set and malware didn’t quite fit what was needed for them.



STIX graph for the ColumbiaChemicals incident

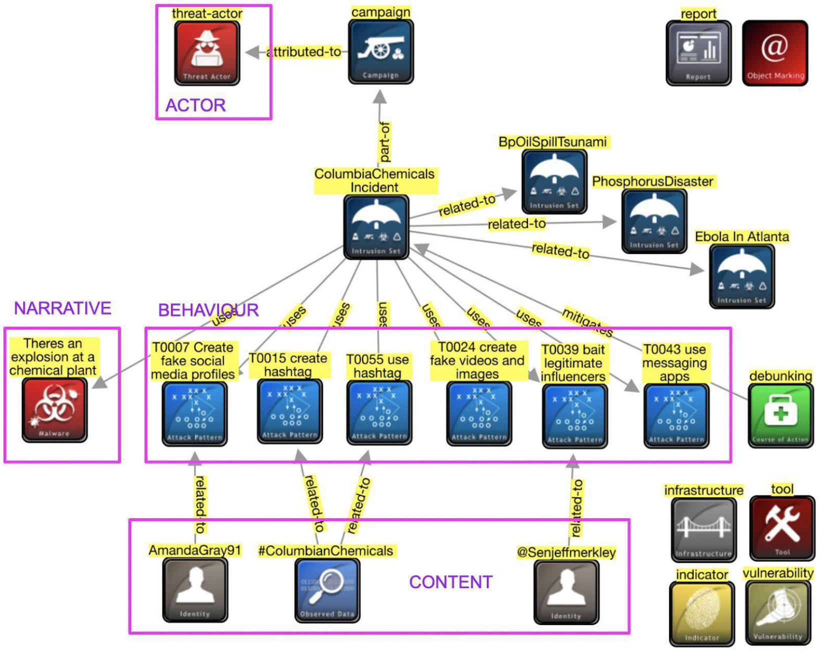
A STIX graph for the Columbia Chemicals incident (a very short-term 2014 incident in the USA) is shown above, with the disinformation pyramid layers (campaign, incident, narrative, artifact) overlaid on it. This helps with thinking about relationships between disinformation layers: a disinformation incident usually belongs to one campaign (although there were many crossover campaigns in 2020, e.g. covid5g, where it was difficult to determine this), but multiple incidents can use the same narratives and artifacts.

There are other component-based disinformation models, notably Camille Francois' ABC "Actor, Behaviour, Content" model and its extension, ABCDE ("Actor, Behaviour, Content, Distribution, Effect"), which adds risk assessment components to assessing an incident.

Actor	*What kinds of actors are involved?* This question can help establish, for example, whether the case involves a foreign state actor
Behavior	*What activities are exhibited?* This inquiry can help establish, for instance, evidence of coordination and inauthenticity
Content	*What kinds of content are being created and distributed?* This line of questioning can help establish, for example, whether the information being deployed is deceptive.
Distribution	*What is the degree to which the content is being distributed?* This question can help establish the incident's reach.
Effect	*What is the overall impact of the case and whom does it affect?* This question can help establish the actual harms and severity of the case.

ABCDE framework for disinformation [Pamment20]

ABC model components and narrative objects are shown in the ColumbiaChemicals diagram below - of note is that these model the disinformation creators' ABC, not the disinformation defenders (e.g. the "debunking" object is outside the Behaviour box).



ABC model components in the ColumbiaChemicals Stix graph

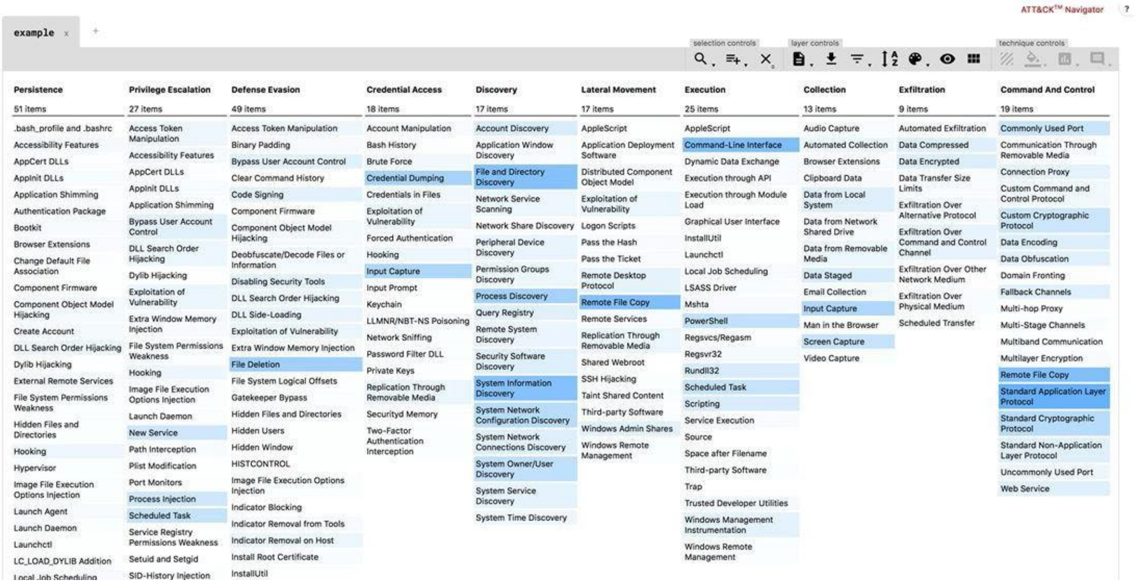
Building a disinformation model based on STIX allows analysts to share and compare information about threat actors, narratives, TTPs, artifacts and other objects in each incident and campaign, using the tools already built for STIX. It also allows disinformation data to flow through the same systems as information security data, making description and countering of hybrid (combinations of disinformation and other infosec methods) easier.

2.4 Behaviour-based Disinformation Models

The infosec community has multiple models that capture the behaviours of incident creators and responders. Several of these models, including MITRE's ATT&CK framework, focus on the techniques, tactics and procedures (TTPs) used by incident creators and responders, where TTPs are the blue ("attack pattern") and green ("course of action") boxes in the STIX diagrams above. Most of MisinfosecWG's effort was on how to adapt these models, and the tools that use them, for disinformation response.

2.4.1 Disinformation Threat Models

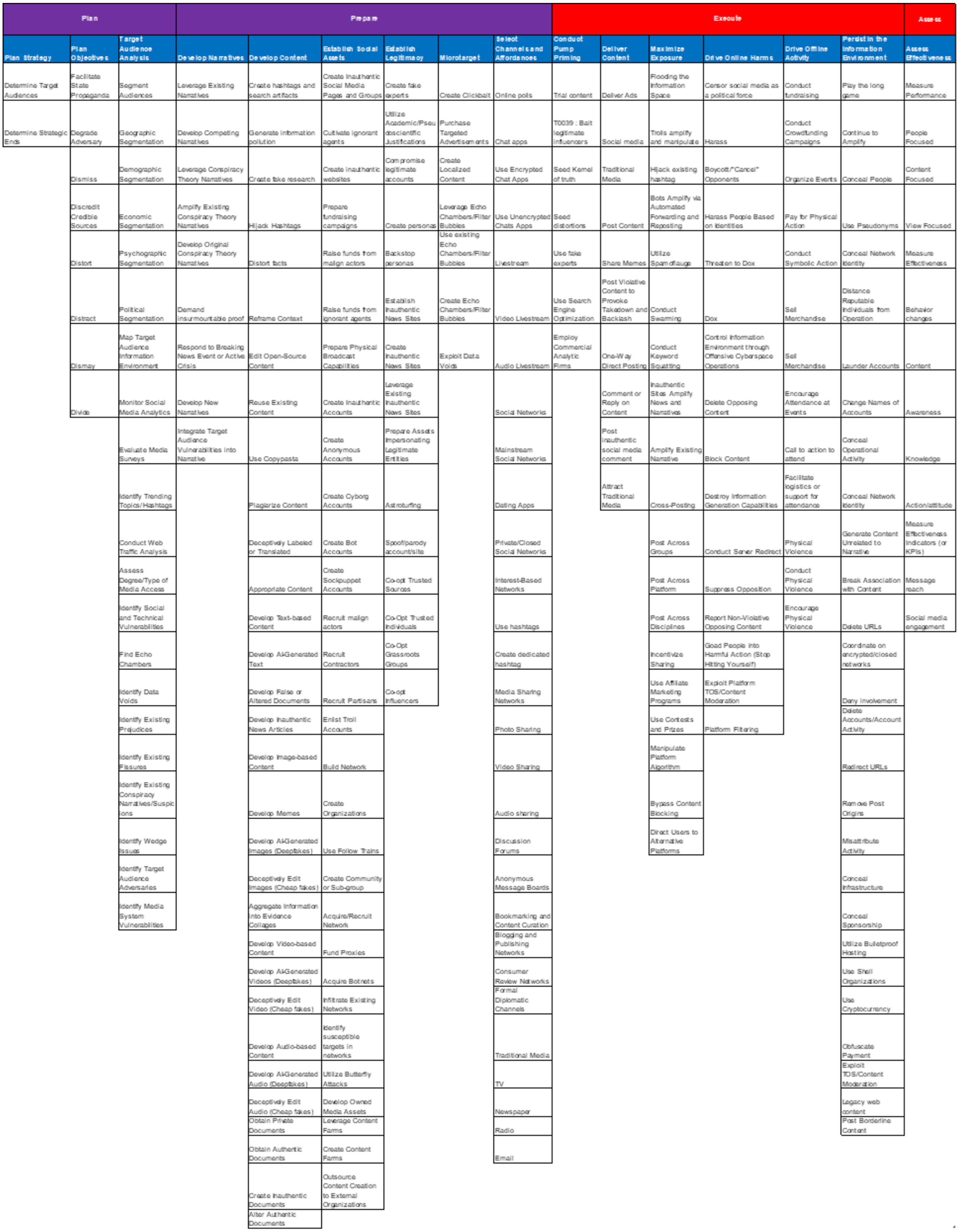
In 2019, MisinfosecWG mined known disinformation incidents for incident creator behaviours, and looked for inspiration and frameworks for disinformation behaviour-based models. The group looked at behaviour-based models from information security, social network analysis, marketing, and adtech before settling on the Cyber Killchain and the ATT&CK model that's based on it, as a base representation for disinformation behaviours.



Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Lateral Movement	Execution	Collection	Exfiltration	Command And Control
51 items	27 items	49 items	18 items	17 items	17 items	25 items	13 items	9 items	19 items
.bash_profile and .bashrc	Access Token Manipulation	Access Token Manipulation	Account Manipulation	Account Discovery	AppleScript	AppleScript	Audio Capture	Automated Exfiltration	Commonly Used Port
Accessibility Features	Accessibility Features	Binary Padding	Bash History	Application Window Discovery	Application Deployment Software	Command-Line Interface	Automated Collection	Data Compressed	Communication Through Removable Media
AppCert DLLs	AppCert DLLs	Bypass User Account Control	Brute Force	File and Directory Discovery	Distributed Component Object Model	Dynamic Data Exchange	Browser Extensions	Data Encrypted	Connection Proxy
AppInit DLLs	AppInit DLLs	Clear Command History	Credential Dumping	Network Service Scanning	Exploitation of Vulnerability	Execution through API	Clipboard Data	Data Transfer Size Limits	Custom Command and Control Protocol
Application Shimming	Application Shimming	Code Signing	Credentials in Files	Network Share Discovery	Logon Scripts	Execution through Module Load	Data from Local System	Exfiltration Over Alternative Protocol	Custom Cryptographic Protocol
Authentication Package	Bypass User Account Control	Component Firmware	Exploitation of Vulnerability	Peripheral Device Discovery	Pass the Hash	Graphical User Interface	Data from Network Shared Drive	Exfiltration Over Command and Control Channel	Data Encoding
Bootkit	DLL Search Order Hijacking	Component Object Model Hijacking	Forced Authentication	Permission Groups Discovery	Pass the Ticket	InstallUtil	Data from Removable Media	Exfiltration Over Other Network Medium	Data Obfuscation
Browser Extensions	Dylib Hijacking	Deobfuscate/Decode Files or Information	Hooking	Process Discovery	Remote Desktop Protocol	Launchctl	Data Staged	Exfiltration Over Physical Medium	Domain Fronting
Change Default File Association	Exploitation of Vulnerability	Disabling Security Tools	Input Capture	Query Registry	Remote File Copy	Local Job Scheduling	Email Collection	Scheduled Transfer	Fallback Channels
Component Firmware	Extra Window Memory Injection	DLL Search Order Hijacking	Input Prompt	Remote System Discovery	Remote Services	LSASS Driver	Input Capture		Multi-hop Proxy
Component Object Model Hijacking	File System Permissions Weakness	DLL Side-Loading	Keychain	Security Software Discovery	Replication Through Removable Media	Mshta	Man in the Browser		Multi-Stage Channels
Create Account	Hooking	Exploitation of Vulnerability	LLMNR/NBT-NS Poisoning	System Information Discovery	Shared Webroot	PowerShell	Screen Capture		Multiband Communication
DLL Search Order Hijacking	Image File Execution Options Injection	Extra Window Memory Injection	Network Sniffing	System Network Configuration Discovery	SSH Hijacking	Regsvcs/Regasm	Video Capture		Multilayer Encryption
Dylib Hijacking	Launch Daemon	File Deletion	Password Filter DLL	System Network Connections Discovery	Taint Shared Content	Regsvr32			Remote File Copy
External Remote Services	New Service	File System Logical Offsets	Private Keys	System Owner/User Discovery	Third-party Software	Rundll32			Standard Application Layer Protocol
File System Permissions Weakness	Path Interception	Gatekeeper Bypass	Replication Through Removable Media	System Service Discovery	Windows Admin Shares	Scheduled Task			Standard Cryptographic Protocol
Hidden Files and Directories	Plist Modification	Hidden Files and Directories	Securityd Memory	System Time Discovery	Windows Remote Management	Scripting			Standard Non-Application Layer Protocol
Hooking	Port Monitors	Hidden Users	Two-Factor Authentication Interception			Service Execution			Uncommonly Used Port
Hypervisor	Process Injection	Hidden Window				Source			Web Service
Image File Execution Options Injection	Scheduled Task	HISTCONTROL				Space after Filename			
Launch Agent	Service Registry Permissions Weakness	Image File Execution Options Injection				Third-party Software			
Launch Daemon	Setuid and Setgid	Indicator Blocking				Trap			
Launchctl	SID-History Injection	Indicator Removal from Tools				Trusted Developer Utilities			
LC_LOAD_DYLIB Addition		Indicator Removal on Host				Windows Management Instrumentation			
Local Job Scheduling		Install Root Certificate				Windows Remote Management			
		InstallUtil							

MITRE ATT&CK framework (Struse)

The model that MisinfosecWG created from this work was what is now the DISARM Framework. The model for the DISARM Red framework (based on the ATT&CK framework) describes common disinformation TTPs across 16 stages of adversary activity, from strategic planning of each incident to evaluating its effectiveness and lessons learned from the deployment, as a feed into later incident plans.



Plan			Prepare						Execute						Assess
Plan Strategy	Plan Objectives	Target Audience Analysis	Develop Narratives	Develop Content	Establish Social Assets	Establish Legitimacy	Microtarget	Select Channels and Affordances	Conduct Pump Priming	Deliver Content	Maximize Exposure	Drive Online Harms	Drive Offline Activity	Persist in the Information Environment	Assess Effectiveness
Determine Target Audiences	Facilitate State Propaganda	Segment Audiences	Leverage Existing Narratives	Create hashtags and search artifacts	Create Inauthentic Social Media Pages and Groups	Create fake experts	Create Clickbait	Online polls	Trial content	Deliver Ads	Flooding the Information Space	Censor social media as a political force	Conduct fundraising	Play the long game	Measure Performance
Determine Strategic Ends	Degrade Adversary	Geographic Segmentation	Develop Competing Narratives	Generate information pollution	Cultivate ignorant agents	Utilize Academic/Pseudoscientific Justifications	Purchase Targeted Advertisements	Chat apps	T0039 : Bait legitimate influencers	Social media	Trolls amplify and manipulate	Harass	Conduct Crowdfunding Campaigns	Continue to Amplify	People Focused
	Dismiss	Demographic Segmentation	Leverage Conspiracy Theory Narratives	Create fake research	Create inauthentic websites	Compromise legitimate accounts	Create Localized Content	Use Encrypted Chat Apps	Seed Kernel of truth	Traditional Media	Hijack existing hashtag	Boycott/"Cancel" Opponents	Organize Events	Conceal People	Content Focused
	Discredit Credible Sources	Economic Segmentation	Amplify Existing Conspiracy Theory Narratives	Hijack Hashtags	Prepare fundraising campaigns	Create personas	Leverage Echo Chambers/Filter Bubbles	Use Unencrypted Chats Apps	Seed distortions	Post Content	Bots Amplify via Automated Forwarding and Reposting	Harass People Based on Identities	Pay for Physical Action	Use Pseudonyms	View Focused
	Distort	Psychographic Segmentation	Develop Original Conspiracy Theory Narratives	Distort facts	Raise funds from malign actors	Backstop personas	Use existing Echo Chambers/Filter Bubbles	Livestream	Use fake experts	Share Memes	Utilize Spamoflauge	Threaten to Dox	Conduct Symbolic Action	Conceal Network Identity	Measure Effectiveness
	Distract	Political Segmentation	Demand insurmountable proof	Reframe Context	Raise funds from ignorant agents	Establish Inauthentic News Sites	Create Echo Chambers/Filter Bubbles	Video Livestream	Use Search Engine Optimization	Post Violative Content to Provoke Takedown and Backlash	Conduct Swarming	Dox	Sell Merchandise	Distance Reputable Individuals from Operation	Behavior changes
	Dismay	Map Target Audience Information Environment	Respond to Breaking News Event or Active Crisis	Edit Open-Source Content	Prepare Physical Broadcast Capabilities	Create Inauthentic News Sites	Exploit Data Voids	Audio Livestream	Employ Commercial Analytic Firms	One-Way Direct Posting	Conduct Keyword Squatting	Control Information Environment through Offensive Cyberspace Operations	Sell Merchandise	Launder Accounts	Content
	Divide	Monitor Social Media Analytics	Develop New Narratives	Reuse Existing Content	Create Inauthentic Accounts	Leverage Existing Inauthentic News Sites		Social Networks		Comment or Reply on Content	Inauthentic Sites Amplify News and Narratives	Delete Opposing Content	Encourage Attendance at Events	Change Names of Accounts	Awareness
		Evaluate Media Surveys	Integrate Target Audience Vulnerabilities into Narrative	Use Copypasta	Create Anonymous Accounts	Prepare Assets Impersonating Legitimate Entities		Mainstream Social Networks		Post inauthentic social media comment	Amplify Existing Narrative	Block Content	Call to action to attend	Conceal Operational Activity	Knowledge
		Identify Trending Topics/Hashtags		Plagiarize Content	Create Cyborg Accounts	Astroturfing		Dating Apps		Attract Traditional Media	Cross-Posting	Destroy Information Generation Capabilities	Facilitate logistics or support for attendance	Conceal Network Identity	Action/attitude
		Conduct Web Traffic Analysis		Deceptively Labeled or Translated	Create Bot Accounts	Spoof/parody account/site		Private/Closed Social Networks			Post Across Groups	Conduct Server Redirect	Physical Violence	Generate Content Unrelated to Narrative	Measure Effectiveness Indicators (or KPIs)
		Assess Degree/Type of Media Access		Appropriate Content	Create Sockpuppet Accounts	Co-opt Trusted Sources		Interest-Based Networks			Post Across Platform	Suppress Opposition	Conduct Physical Violence	Break Association with Content	Message reach
		Identify Social and Technical Vulnerabilities		Develop Text-based Content	Recruit malign actors	Co-Opt Trusted Individuals		Use hashtags			Post Across Disciplines	Report Non-Violative Opposing Content	Encourage Physical Violence	Delete URLs	Social media engagement
		Find Echo Chambers		Develop AI-Generated Text	Recruit Contractors	Co-Opt Grassroots Groups		Create dedicated hashtag			Incentivize Sharing	Goad People into Harmful Action (Stop Hitting Yourself)		Coordinate on encrypted/closed networks	
		Identify Data Voids		Develop False or Altered Documents	Recruit Partisans	Co-opt Influencers		Media Sharing Networks			Use Affiliate Marketing Programs	Exploit Platform TOS/Content Moderation		Deny involvement	
		Identify Existing Prejudices		Develop Inauthentic News Articles	Enlist Troll Accounts			Photo Sharing			Use Contests and Prizes	Platform Filtering		Delete Accounts/Account Activity	
		Identify Existing Fissures		Develop Image-based Content	Build Network			Video Sharing			Manipulate Platform Algorithm			Redirect URLs	
		Identify Existing Conspiracy Narratives/Suspicions		Develop Memes	Create Organizations			Audio sharing			Bypass Content Blocking			Remove Post Origins	
		Identify Wedge Issues		Develop AI-Generated Images (Deepfakes)	Use Follow Trains			Discussion Forums			Direct Users to Alternative Platforms			Misattribute Activity	
		Identify Target Audience Adversaries		Deceptively Edit Images (Cheap fakes)	Create Community or Sub-group			Anonymous Message Boards						Conceal Infrastructure	
		Identify Media System Vulnerabilities		Aggregate Information into Evidence Collages	Acquire/Recruit Network			Bookmarking and Content Curation						Conceal Sponsorship	
				Develop Video-based Content	Fund Proxies			Blogging and Publishing Networks						Utilize Bulletproof Hosting	
				Develop AI-Generated Videos (Deepfakes)	Acquire Botnets			Consumer Review Networks						Use Shell Organizations	
				Deceptively Edit Video (Cheap fakes)	Infiltrate Existing Networks			Formal Diplomatic Channels						Use Cryptocurrency	
				Develop Audio-based Content	Identify susceptible targets in networks			Traditional Media						Obfuscate Payment	
				Develop AI-Generated Audio (Deepfakes)	Utilize Butterfly Attacks			TV						Exploit TOS/Content Moderation	
				Deceptively Edit Audio (Cheap fakes)	Develop Owned Media Assets			Newspaper						Legacy web content	
				Obtain Private Documents	Leverage Content Farms			Radio						Post Borderline Content	
				Obtain Authentic Documents	Create Content Farms			Email							
				Create Inauthentic Documents	Outsource Content Creation to External Organizations										
				Alter Authentic Documents											

The DISARM Red Framework

The DISARM framework has three main component types:

- Phases (e.g. “Plan”)
- Tactic Stages (e.g. “Plan Strategy”): the set of top-level adversary goals that are needed to complete a successful attack.
- Tactics, techniques and procedures (TTPs, e.g. “Determine Target Audiences”): the means by which incident creators meet each tactic goal.

With a behaviour-based framework, we can start to record and recall previous countermeasures to reused techniques, and find and exploit weaknesses and gaps in the adversary’s operations, in the same way we exploit adversary weaknesses in gaps in other situation pictures, including those in cybersecurity.

2.4.2 Disinformation Response Models

TTPs that model the behaviour of disinformation creators are one half of the behaviour-based models that we need for disinformation defence. In late 2019, MisinfosecWG extended its work to model the countermeasure and mitigation actions available to disinformation defenders.

Information security already has models for this: the course of action objects seen in STIX above. These are usually shown in a “Courses of Action Matrix” - a grid where tactic stages are plotted against different categories of countermeasure.

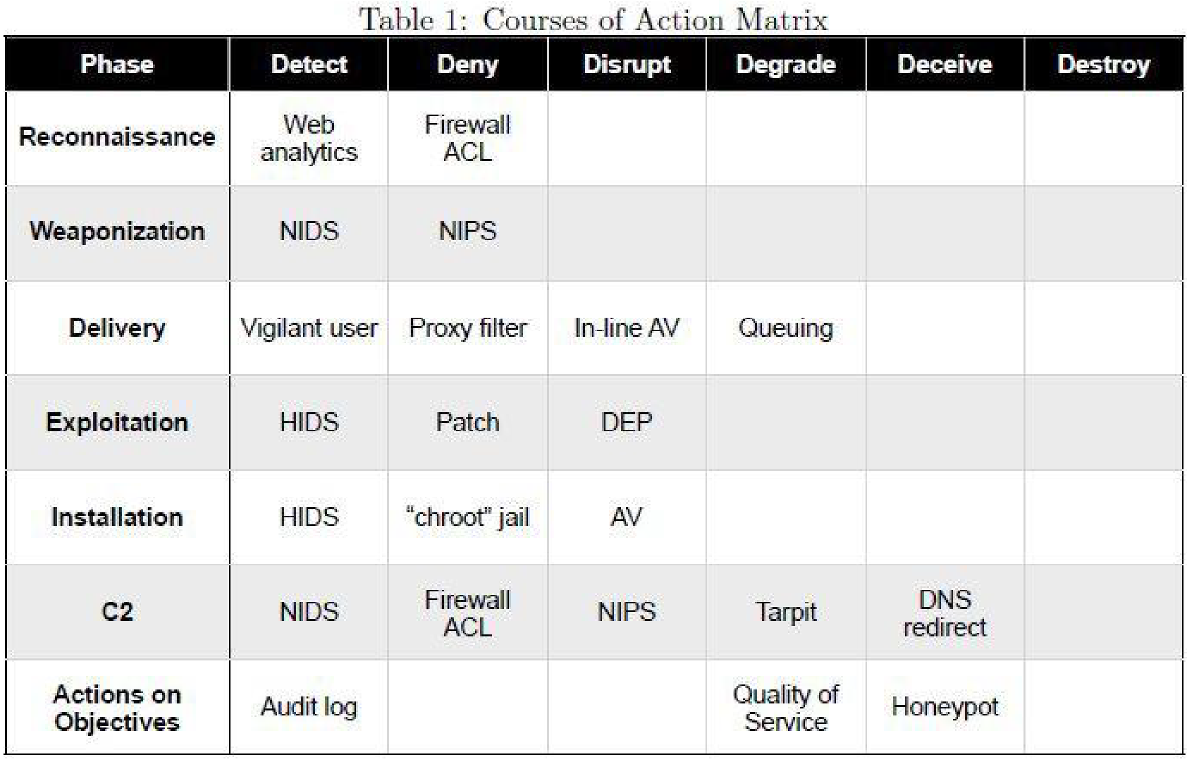


Table 1: Courses of Action Matrix

Phase	Detect	Deny	Disrupt	Degrade	Deceive	Destroy
Reconnaissance	Web analytics	Firewall ACL				
Weaponization	NIDS	NIPS				
Delivery	Vigilant user	Proxy filter	In-line AV	Queuing		
Exploitation	HIDS	Patch	DEP			
Installation	HIDS	"chroot" jail	AV			
C2	NIDS	Firewall ACL	NIPS	Tarpit	DNS redirect	
Actions on Objectives	Audit log			Quality of Service	Honeypot	

Cyber Killchain Courses of Action Matrix

A courses of action matrix for the cyber killchain (the model we based DISARM on) is shown above. Down the left side we have the seven cyber killchain tactic stages. Along the top we have six types of countermeasure or mitigation effect. Each grid square contains suggested actions that could create that effect on that tactic stage.

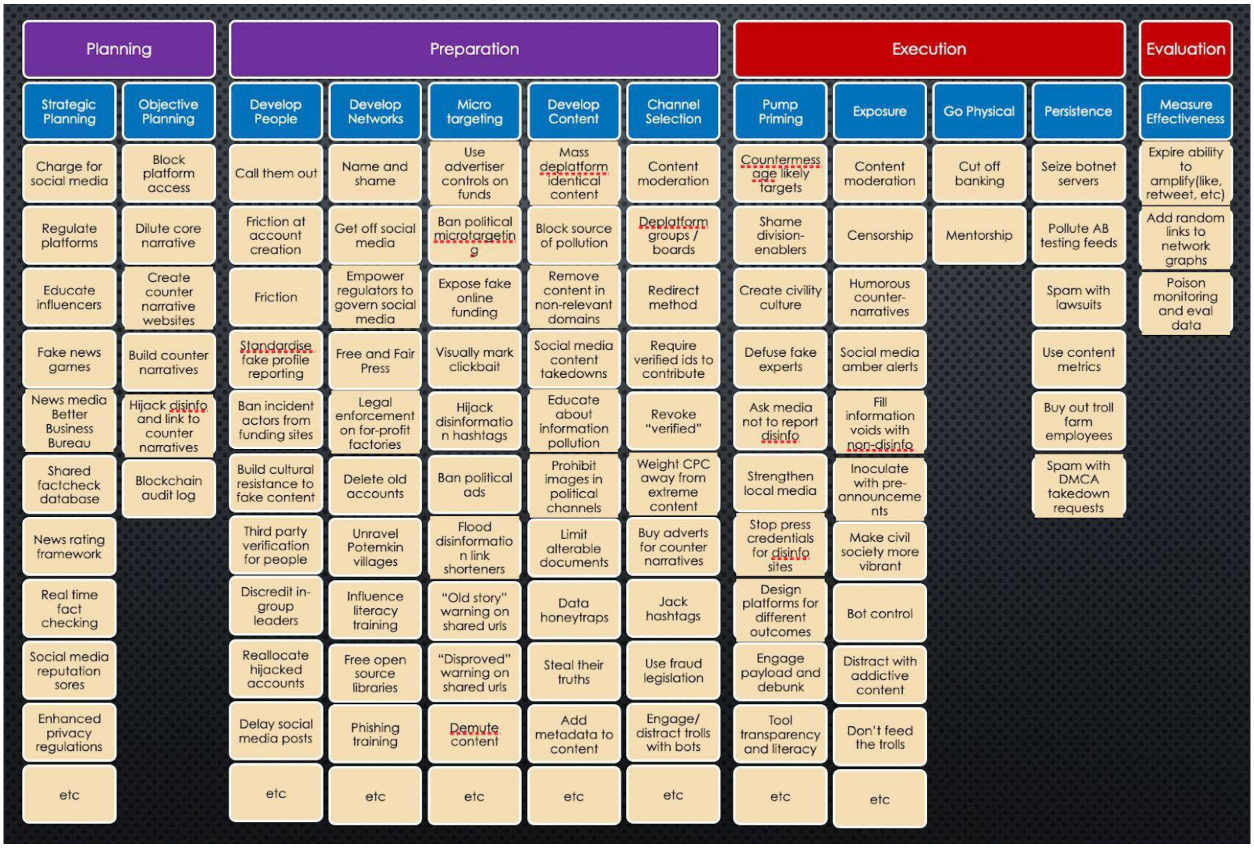
MisinfosecWG examined the disinformation solution space, considering the tools and techniques that existed and might be needed in it, then ran a Courses of Action generating exercise for the tactic stages, producing countermeasures and mitigations organised by countermeasure type, tactic stage and TTP. This formed a labeled list of disinformation creator TTPs that the CogSecCollab team extended to include the resources needed to deploy each countermeasure, and example playbooks for several counters.

	D2 Deny	D3 Disrupt	D4 Degrade	D5 Deceive	D6 Destroy	D7 Deter	TOTALS
TA01 Strategic Planning	11	6	7			4	28
TA02 Objective Planning		5					5
TA03 Develop People	10	7	1	1	1	1	21
TA04 Develop Networks	11	3	3		1		18
TA05 Microtargeting	2	5					7
TA06 Develop Content	13	8	5	2		5	33
TA07 Channel Selection	7	7	3	1			18

TA08 Pump Priming	7	3	2			3	15
TA09 Exposure	3	14	2				19
TA10 Go Physical	1					1	2
TA11 Persistence	1	6	6				13
TA12 Measure Effectiveness		1	2				3

Original counter TTP counts for tactic stages and counter types

This exercise produced more than 100 AMITT countermeasure TTPs that were listed alongside the incident creator TTPs that they mitigate or counter.



Planning		Preparation					Execution				Evaluation
Strategic Planning	Objective Planning	Develop People	Develop Networks	Micro targeting	Develop Content	Channel Selection	Pump Priming	Exposure	Go Physical	Persistence	Measure Effectiveness
Charge for social media	Block platform access	Call them out	Name and shame	Use advertiser controls on funds	Mass deplatform identical content	Content moderation	Countermessage likely targets	Content moderation	Cut off banking	Seize botnet servers	Expire ability to amplify(like, retweet, etc)
Regulate platforms	Dilute core narrative	Friction at account creation	Get off social media	Ban political microtargeting	Block source of pollution	Deplatform groups / boards	Shame division-enablers	Censorship	Mentorship	Pollute AB testing feeds	Add random links to network graphs
Educate influencers	Create counter narrative websites	Friction	Empower regulators to govern social media	Expose fake online funding	Remove content in non-relevant domains	Redirect method	Create civility culture	Humorous counter-narratives		Spam with lawsuits	Poison monitoring and eval data
Fake news games	Build counter narratives	Standardise fake profile reporting	Free and Fair Press	Visually mark clickbait	Social media content takedowns	Require verified ids to contribute	Defuse fake experts	Social media amber alerts		Use content metrics	
News media Better Business Bureau	Hijack disinfo and link to counter narratives	Ban incident actors from funding sites	Legal enforcement on for-profit factories	Hijack disinformation hashtags	Educate about information pollution	Revoke "verified"	Ask media not to report disinfo	Fill information voids with non-disinfo		Buy out troll farm employees	
Shared factcheck database	Blockchain audit log	Build cultural resistance to fake content	Delete old accounts	Ban political ads	Prohibit images in political channels	Weight CPC away from extreme content	Strengthen local media	Inoculate with pre-announcements		Spam with DMCA takedown requests	
News rating framework		Third party verification for people	Unravel Potemkin villages	Flood disinformation link shorteners	Limit alterable documents	Buy adverts for counter narratives	Stop press credentials for disinfo sites	Make civil society more vibrant			
Real time fact checking		Discredit in-group leaders	Influence literacy training	"Old story" warning on shared urls	Data honeytraps	Jack hashtags	Design platforms for different outcomes	Bot control			
Social media reputation sores		Reallocate hijacked accounts	Free open source libraries	"Disproved" warning on shared urls	Steal their truths	Use fraud legislation	Engage payload and debunk	Distract with addictive content			
Enhanced privacy regulations		Delay social media posts	Phishing training	Demute content	Add metadata to content	Engage/ distract trolls with bots	Tool transparency and literacy	Don't feed the trolls			
etc		etc	etc	etc	etc	etc	etc	etc			

Original AMITT countermeasures TTP diagram

The DISARM countermeasures TTP diagram is currently larger than the incident TTPs diagram, as we have worked to integrate SP!CE and clean it up and place techniques into the right stages.

2.4.3 Multiplayer game models

With the DISARM STIX, DISARM framework TTPs and DISARM countermeasure TTPs in place, it’s possible to start modelling disinformation ecosystems as simulations or games in which multiple players compete for limited resources including narratives, attention and time. Threet designed models that organize resources, so multiplayer, multi-viewpoint games and simulations could be designed using the existing DISARM TTPs and objects. Another multi-player view of the disinformation solution space is as a human space, in which narratives compete for dominance (e.g “narrative warfare”). Human communication is generally at the level of stories, or narration: we tell each other stories about the world, as sentences, image sequences, or memes. Narratives are the stories that each person and community bases their sense of self, their belonging to different groups (“in-groups”), and exclusion of others (“out-groups”) on. Narratives are typically personal, emotionally-charged, deeply-entrenched and difficult to shift directly. In this space, it becomes important to track narratives and their components (e.g. memes, stories, sentiments) and disrupt them not by countering them directly with ‘facts’, but with ‘information aikido’: it’s easier to redirect an angry mob to a different house than it is to disband them. Narrative warfare is a growing field, and its techniques are a useful component in countering disinformation. Using Natural Language Processing techniques like topic modelling and gisting to track narratives from disinformation actors, and highlighting narratives to potential target audiences have also proved useful. DISARM models don’t explicitly include narrative warfare or machine learning models, although these have been built and studied independently by the DISARM teams.

2.5 Work in Progress

2.5.1 Disinformation Risk Modelling

Disinformation is a form of digital harm, alongside hate speech, cyber bullying, fraud, spam and other activities that potentially damage individuals, groups etc. digital harms can be managed as risks, where a risk is defined as a combination of severity, likelihood and target. SJ is working separately on disinformation risk models – these are useful in triaging (deciding which incidents to put response resources onto) misinformation, disinformation and threat actors.

In 2020, CogSecCollab used basic risk models to triage incidents coming into the CTI League and other deployments. These could be extended using the “DE” part of the ABCDE model, to give risk assessment and triage at other levels of the disinformation pyramid.

2.5.2 Disinformation Taxonomies

DISARM object types are not sufficient to completely describe a disinformation incident. DISARM STIX is missing categories for each of its object types.

For instance, DISARM STIX contains “threat actor”, but doesn’t have a set of labels for possible types of threat actor – geopolitically motivated, financially motivated etc, to make it easier for a user or information recipient to determine if a new actor is of interest to them.

Existing taxonomies of disinformation object types include DFRlab’s dichotomies of disinformation, which are designed for strategic analysis of disinformation actors, campaigns and incidents. In 2019, CogSecCollab worked with NATO to produce a taxonomy based on DFRlab’s taxonomy, but better suited for fast-paced tactical use. Agile, and the limits of standards-based approaches

At this stage, older infosec people are probably shaking their heads and muttering something about stamp collecting and bingo cards. We get that. We know that defending against a truly agile adversary isn’t a game of lookup, and as fast as we design and build counters, our counterparts will build counters to the counters, new techniques, new adaptations of existing techniques etc. Adversary tactics are moving quickly in this arena (for instance, the types of tool changes already seen in the related field of Mlsec), so tools and counter tactics are likely to change but the basic problems won’t.

But that’s only part of the game. Most of the time people get lazy, or get into a rut — they reuse techniques and tools, or it’s too expensive to keep moving. It makes sense to build descriptions like this that we can adapt over time. It also helps us spot when we’re outside the frame.

There is no one, magic, response to misinformation. Misinformation mitigation, like disease control, is a whole-system response. All the tools mentioned above are intended for use by threat intelligence teams, often working in near-real-time from Security Operations Centers and their equivalents.

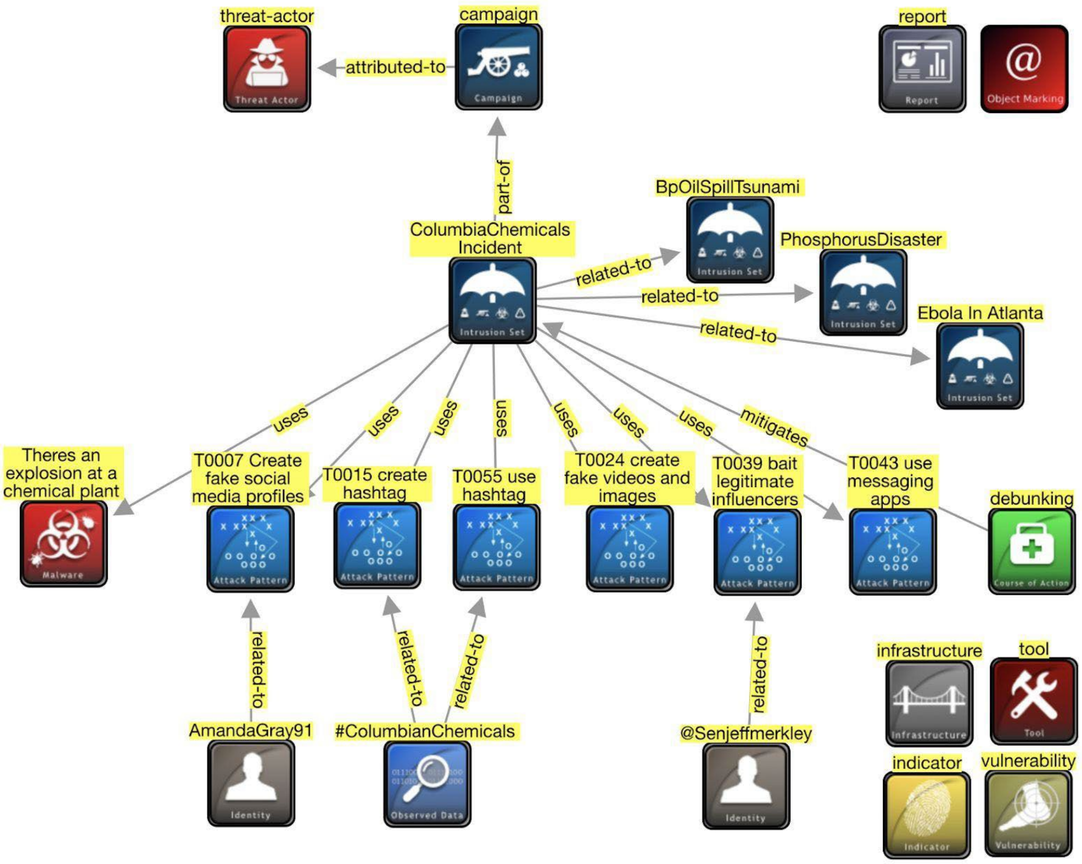
Sometimes you just respond, but it helps to do this from a position of knowledge, shared communication and respect for the potential risks to actors, organisations, narratives and other components of the information ecosystem we’re working in. MisinfosecWG looked at Adam Shostack’s slides on threat modelling in 2019, and specifically at the differences between slower “waterfall, V” threat models (STRIDE, kill chain etc), and faster-reacting “agile” and lean threat models, where agile is rapidly iterating over solutions in a known problem space, and lean is iterations on both the problem and solution spaces. This is one of the considerations when designing tactical disinformation response: we still need the slower, deliberative work that gives labels and lists defences and counters for common threats (the “phishing” etc equivalents of cognitive security), but we also need that rapid response to things previously unseen that keeps white-hat hackers glued to their screens for hours. There’s more about this in CogSecCollab’s writings on creating and operating disinformation Security

Operations Centres.

2.6 Further Reading

- WWW 2019 AMITT paper; summary of AMITT WWW paper

DISARM STIX Design and Philosophy



STIX graph for the Columbia Chemicals incident

STIX is a data standard used to share information between threat intelligence organisations like ISACs. It's a rich language that describes threat objects and the relationships between them, is extensible, used by existing threat intelligence sharing communities \(ISACs, ISAOs etc\) so we'd be patching into an existing sharing system. It's also supported by and integrates well with existing community-supported, open-source tools. STIX translates well for disinformation use.

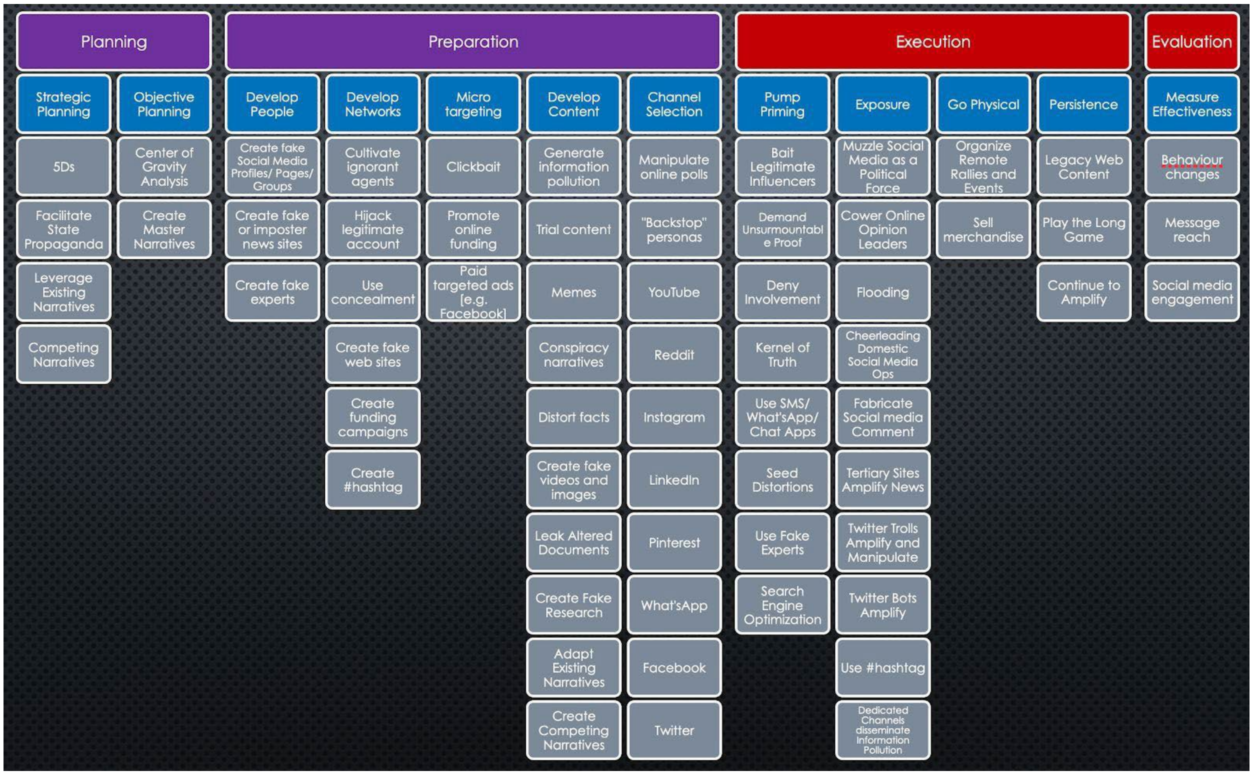
Disinformation STIX	Description	Level	Infosec STIX
Report	communication to other responders	Communication	Report
Campaign	Longer attacks (Russia's interference in the 2016 US elections is a "campaign")	Strategy	Campaign
Incident	Shorter-duration attacks, often part of a campaign	Strategy	Intrusion Set
Course of Action	Response	Strategy	Course of Action
Identity	Actor (individual, group, organization etc): creator, responder, target, useful idiot etc.	Strategy	Identity
Threat actor	Incident creator	Strategy	Threat Actor
Attack pattern	Technique used in incident (see framework for examples)	TTP	Attack pattern
Narrative	Malicious narrative (story, meme)	TTP	Malware
Tool	bot software, APIs, marketing tools	TTP	Tool
Observed Data	artefacts like messages, user accounts, etc	Artefact	Observed Data
Indicator	posting rates, follow rates etc	Artefact	Indicator
Vulnerability	Cognitive biases, community structural weakness etc	Vulnerability	Vulnerability

Mappings Between infosec STIX and cogsec STIX

We added two objects to STIX for disinformation: incident, and narrative, and didn't need to change anything else. We use custom objects to represent these fields and be OpenCTI compliant.

AMITT, before being renamed to DISARM, was available as a STIX 2.0 bundle, from https://github.com/cogsec-collaborative/amitt_cti . With STIX 2.1 an incident object has become available which we have migrated to. The current STIX 2.1 bundle can be found at https://github.com/DISARMFoundation/DISARM-STIX2.

3 DISARM Framework Design and Philosophy

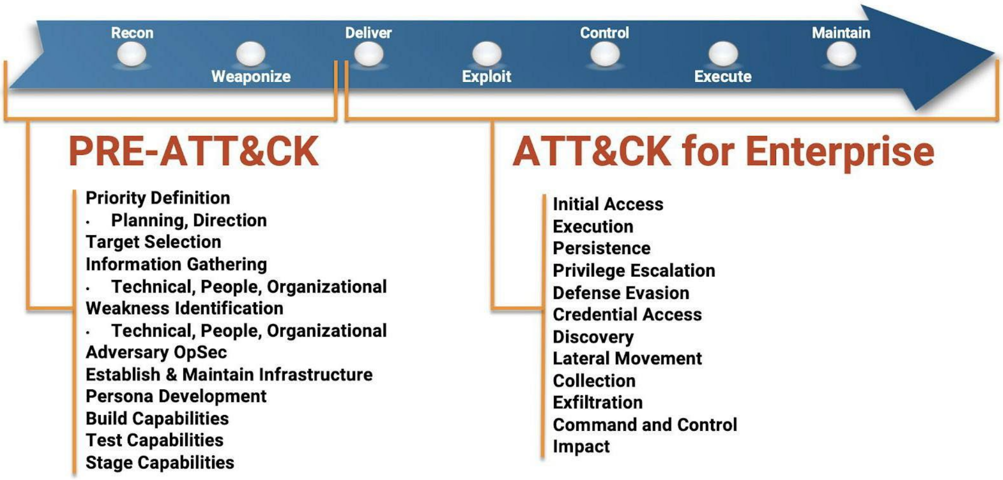


Planning		Preparation					Execution				Evaluation
Strategic Planning	Objective Planning	Develop People	Develop Networks	Micro targeting	Develop Content	Channel Selection	Pump Priming	Exposure	Go Physical	Persistence	Measure Effectiveness
5Ds	Center of Gravity Analysis	Create fake Social Media Profiles/ Pages/ Groups	Cultivate ignorant agents	Clickbait	Generate information pollution	Manipulate online polls	Bait Legitimate Influencers	Muzzle Social Media as a Political Force	Organize Remote Rallies and Events	Legacy Web Content	Behaviour changes
Facilitate State Propaganda	Create Master Narratives	Create fake or imposter news sites	Hijack legitimate account	Promote online funding	Trial content	"Backstop" personas	Demand Unsurmountable Proof	Cower Online Opinion Leaders	Sell merchandise	Play the Long Game	Message reach
Leverage Existing Narratives		Create fake experts	Use concealment	Paid targeted ads [e.g. Facebook]	Memes	YouTube	Deny Involvement	Flooding		Continue to Amplify	Social media engagement
Competing Narratives			Create fake web sites		Conspiracy narratives	Reddit	Kernel of Truth	Cheerleading Domestic Social Media Ops			
			Create funding campaigns		Distort facts	Instagram	Use SMS/ What'sApp/ Chat Apps	Fabricate Social media Comment			
			Create #hashtag		Create fake videos and images	LinkedIn	Seed Distortions	Tertiary Sites Amplify News			
					Leak Altered Documents	Pinterest	Use Fake Experts	Twitter Trolls Amplify and Manipulate			
					Create Fake Research	What'sApp	Search Engine Optimization	Twitter Bots Amplify			
					Adapt Existing Narratives	Facebook		Use #hashtag			
					Create Competing Narratives	Twitter		Dedicated Channels disseminate Information Pollution			

The original AMITT Framework

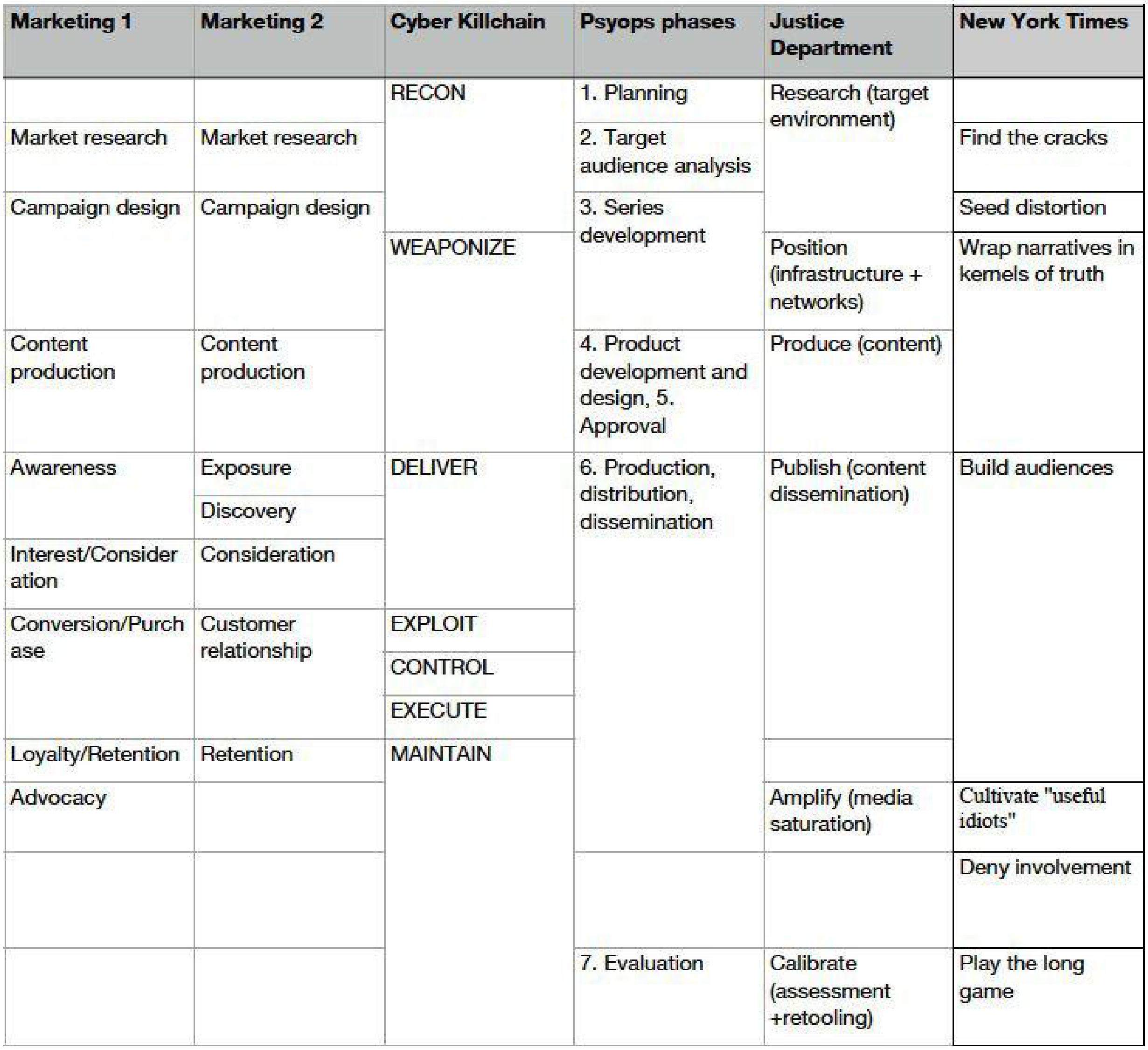
The original AMITT framework was created from a need to describe disinformation behaviours in consistent, concise ways that could allow rapid sharing of information across responding groups. The original framework has 12 stages. The newer DISARM framework is an evolution of the original AMITT framework. It has 16 stages (or tactics).

3.1 Seeding the Model



Top: Cyber Killchain stages, Bottom: ATT&CK framework stages (MITRE intro to ATT&CK)

MisinfosecWG mapped the other main models it was considering for the AMITT framework onto the cyber killchain, to ensure it missed as little as possible from them.



Marketing 1	Marketing 2	Cyber Killchain	Psyops phases	Justice Department	New York Times
		RECON	1. Planning	Research (target environment)	
Market research	Market research		2. Target audience analysis		Find the cracks
Campaign design	Campaign design		3. Series development		Seed distortion
		WEAPONIZE		Position (infrastructure + networks)	Wrap narratives in kernels of truth
Content production	Content production		4. Product development and design, 5. Approval	Produce (content)	
Awareness	Exposure	DELIVER	6. Production, distribution, dissemination	Publish (content dissemination)	Build audiences
	Discovery				
Interest/Consideration	Consideration				
Conversion/Purchase	Customer relationship	EXPLOIT			
		CONTROL			
		EXECUTE			
Loyalty/Retention	Retention	MAINTAIN			
Advocacy				Amplify (media saturation)	Cultivate "useful idiots"
					Deny involvement
			7. Evaluation	Calibrate (assessment +retooling)	Play the long game

Comparison between cyber killchain, marketing, psyops and other potential models

3.2 Organising the DISARM Framework

DISARM's phases are grouped into activities typically performed before a disinformation incident become publicly visible, and those after incident artifacts are widely visible online. The phases before public visibility are termed "left of boom" and are colored purple; those after are "right of boom" and are colored red (this is an old explosive disposal term used in some infosec models).

Like ATT&CK, DISARM’s tactic stages are listed sequentially from left to right - the further left that a tactic is on the DISARM diagram, the earlier that it's likely to be met by an incident creator. In DISARM, tactics are also grouped into four phases: planning, preparation, execution and evaluation; phases are used to evaluate things like the potential for both attacker and defender automations. Each DISARM TTP description includes examples of its use, defender TTPs that could be used to counter or mitigate it, and indicators that could be used to detect it.

Sub-techniques are lower-level, very specific techniques. They are shown on the main DISARM framework diagram with a numerical suffix such as .001, .002 etc. .

3.2.1 Tactic Phases

Tactic stage	**Threat actor is trying to...**	**Techniques**	**KillChain Phase**
Plan Strategy	Define the desired end state, i.e. the set of required conditions that defines achievement of all objectives.	2	Recon
Plan Objectives	Set clearly defined, measurable, and achievable objectives. Achieving objectives ties execution of tactical tasks to reaching the desired end state.	8	Recon
Target Audience Analysis	Identify and analyze the target audience including audience member locations, political affiliations, financial situations, and other attributes that an influence operation may incorporate into its messaging strategy.	21	Recon
Develop Narratives	Promote stories that will gain traction among the target audience and gain narrative dominance.	9	Weaponize
Develop content	Create or acquire text, images, and other content used in incident	31	Weaponize
Establish Social Assets	Establish information assets such as social media accounts, operation personnel, and organizations, including directly and indirectly managed assets.	30	Weaponize
Establish Legitimacy	Establish special information assets that create trust	15	Weaponize

Microtarget	Target very specific populations of people	7	Weaponize
Select Channels and Affordances	Select platforms and affordances to maximize an influence operation’s ability to reach its target audience.	29	Weaponize
Conduct Pump priming	Release content on a targeted small scale, prior to general release, including releasing seed.	7	Deliver
Deliver Content	Release content to general public or larger population	10	Deliver
Maximize Exposure	Maximize exposure of the target audience to incident/campaign content via flooding, amplifying, and cross-posting.	19	Execute
Drive Online Harms	Take actions which harm opponents in online spaces through harassment, suppression, releasing private information, and offensive cyberspace operations.	16	Execute
Drive Offline Activity	Encourage users to engage in the physical information space or offline world. This may include operation-aligned rallies or protests, radio, newspaper, or billboards.	13	Execute
Persist in the Information Environment	Keep incident 'alive' beyond the incident creators' efforts, by taking measures that allow an operation to maintain its presence and avoid takedown by an external entity.	28	Maintain
Assess Effectiveness	Measure the effectiveness of action, for use in future plans	13	Maintain

3.3 Further Reading

ATT&CK models

- MITRE, “ATTACK_Design_and_Philosophy”, 2020
- MITRE, “Getting started with ATT&CK”, October 2019

DISARM

- SJ Terp, “Misinformation has stages”, Misinfocon 2019

4 DISARM Countermeasures Design and Philosophy

The DISARM countermeasures framework was created from a need to move from "admiring the problem", to actively responding to and mitigating for disinformation in as close to real time as sensibly possible.

This section looks at existing and potential disinformation countermeasures and mitigations. It's part of a series of work on how information security principles and practices can be used to improve our understanding of and responses to disinformation campaigns and incidents. This work has resulted in a collection of countermeasures that we are calling the DISARM Blue framework.

4.1 Finding Countermeasures

4.1.1 Introduction

But right now, it's still part of the "admiring the problem" collection of misinformation tools to be truly useful, DISARM needs to contain not just the breakdown of what the blue team thinks the red team is doing, but also what the blue team might be able to do about it. Colloquially speaking, we're talking about countermeasures here.

There are several ways to go about finding countermeasures to any action:

- Look at counters that already exist. We've logged a few already in the DISARM repo, against specific techniques — for example, we listed a set of counters from the Macron election team as part of incident I00022.
- Look at DISARM's parent models - the ATT&CK framework, the psyops model, marketing models etc - and see how they modelled and described counters (e.g look at the mitigations for ATT&CK T1193 Spear phishing).
- Pick a specific tactic, technique or procedure and brainstorm how to counter it — the MisinfosecWG did this as part of their Atlanta retreat, describing potential new counters for two of the techniques on the DISARM framework.
- Wargame red v blue in a 'safe' environment, and capture the counters that people start using. The Rootzbook exercise that Win and Aaron ran at Defcon AI Village was a good start on this, and holds promise as a training and learning environment.
- Run a machine learning algorithm to generate random countermeasures until one starts looking more sensible/effective than the others. Well, perhaps not, but there's likely to be some measure of automation in counters eventually…

MisinfosecWG mapped out misinformation responses, e.g.

- At the technique level — T0025 leak altered documents was countered in France during the Macron election.
- At the tactic level — we can create a courses of action matrix that lists ways to detect, deny, disrupt, degrade, deceive or destroy activities in each tactic stage.
- At the procedure level — we can look at sequences of responses that may be more effective than individual responses in isolation.

4.1.2 Searching for Countermeasures

Searching for disinformation resources at the end of 2019 is much easier than in previous years. Major lists of projects, reports and groups that yielded existing countermeasures included

- Oxford Internet Institute's computational propaganda project's resource finder https://navigator.oii.ox.ac.uk/resources/?resource_filter%5Bsubject%5D%5B%5D=disinf ormation-counter-strategies#
- Rand.org's reports on disinformation (e.g. [Rand2740])
- Scott Yate's CCC lists of projects, and the Credibility Coalition's navigator

Many other groups (CMU etc) are creating their own lists, making this a great time to hunt for specific counters.

4.1.3 Known Countermeasures

There are many published "solutions" to disinformation attacks. While useful, it's foolish to consider any of these the "silver bullet" that solves a disinformation problem; they often address smaller pieces of an attack, or are intractable or don't scale. Disinformation campaigns are whole-system attacks: to solve them we need to look at whole-system solutions: this is more of a "thousand bullet" solution than a single-bullet one. Some components in the current counter landscape are:

- Detecting artificial amplification. Many disinformation campaigns rely on signal amplification, either through 'useful idiots' or by raising message visibility using non-human traffic ('bots' and 'botnets'). Databases of known online bad actors and state-sponsored actors, with data from pages and social media feeds from these actors have proven useful places to look for emerging narratives and links to new actors and artefacts. Tracking bots and botnets has become more difficult as adversaries adapt to detection techniques (both from disinformation detection but also from adjacent domains including mitigating advertising click fraud) and trade message reach for keeping valuable networks online, but there is still value in simple bot/botnet detection techniques including analysis of similarities across accounts linked by topic, hashtags, retweets, references etc, and time-series analysis to check for sleep/wake patterns, activity correlations etc, especially with adversaries new to this space.

- Detecting related artifacts. Disinformation campaigns rarely use one account, platform, account network or domain, and financially-motivated campaigns sometimes run sets of sites with wildly different topics or demographic/country targets. Most work on this isn't tool-based; it's digital forensics, tracking artifacts like tag IDs, domain registrations and reused/linked content across the internet using OSINT tools (Bellingcat and DigitalSherlocks both publish good examples of this work).
- Mitigating artificial amplification. Most current work on this is platform takedowns or "shadow-banning" of known bot, botnet, troll or other artificial amplification social media accounts. Related work includes removal of online advertising and product revenue from domains that are part of financially-motivated disinformation campaigns.
- Resilience against adversarial narratives. It's preferable to remove a disinformation campaign before it reaches the general population, but if it does, building resilience to disinformation campaigns in the form of awareness of techniques, critical reasoning skills etc is useful. Most population resilience counters are in the form of education - either at school level or through information campaigns like the US State Department's War on Pineapple posters. More active population resilience measures include the Baltic Elves volunteer groups posting disclaimers and counter-narratives to Russian disinformation in their countries.

Education is an important counter, but won't be enough on its own. Other counters that are likely to be trialled with it include:

- Tracking data providence to protect against context attacks (digitally sign media and metadata in a way that media includes the original URL in which it was published and private key is that of the original author/publisher)
- Forcing products altered by AI/ML to notify their users (e.g. there was an effort to force Google's very believable AI voice assistant to announce it was an AI before it could talk to customers)
- Requiring legitimate news media to label editorials as such
- Participating in the Cognitive Security Information Sharing and Analysis Organization (ISAO)
- Forcing paid political ads on the Internet to follow the same rules as paid

political advertisements on television

- Baltic community models, e.g. Baltic “Elves” teamed with local media etc

Jonathan Stray’s paper “Institutional Counter-disinformation Strategies in a Networked Democracy” is a good primer on counters available on a national level.

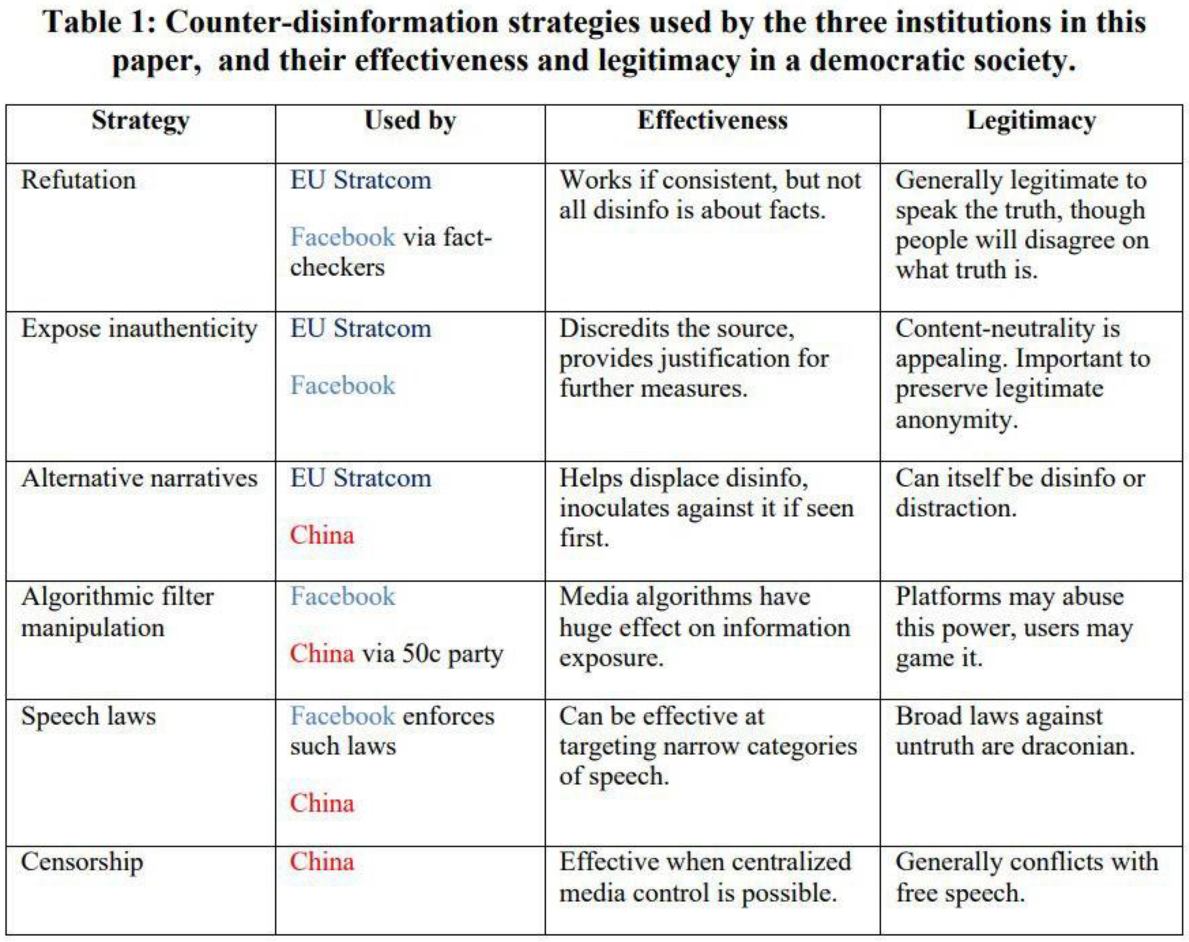


Table 1: Counter-disinformation strategies used by the three institutions in this paper, and their effectiveness and legitimacy in a democratic society.

Strategy	Used by	Effectiveness	Legitimacy
Refutation	EU Stratcom Facebook via fact-checkers	Works if consistent, but not all disinfo is about facts.	Generally legitimate to speak the truth, though people will disagree on what truth is.
Expose inauthenticity	EU Stratcom Facebook	Discredits the source, provides justification for further measures.	Content-neutrality is appealing. Important to preserve legitimate anonymity.
Alternative narratives	EU Stratcom China	Helps displace disinfo, inoculates against it if seen first.	Can itself be disinfo or distraction.
Algorithmic filter manipulation	Facebook China via 50c party	Media algorithms have huge effect on information exposure.	Platforms may abuse this power, users may game it.
Speech laws	Facebook enforces such laws China	Can be effective at targeting narrow categories of speech.	Broad laws against untruth are draconian.
Censorship	China	Effective when centralized media control is possible.	Generally conflicts with free speech.

"A taxonomy of tactics" [Stray19]

4.1.4 Countering DISARM components

Work on DISARM used existing information security models (e.g. cyber kill chain, ATT&CK) to model disinformation incidents as collections of tactics, techniques and procedures (TTPs). One way to look at counters is to look at that breakdown and find or devise responses and mitigations to each TTP. At the tactic level, this gives us a Courses of Action matrix (COA), with the tactic stages listed on one axis, and types of response - eg. (Deny, Disrupt, Degrade, Destroy) - on the other, At the technique level, this gives us a way to discuss mitigations for techniques (e.g. the use of botnets) that we see repeatedly in disinformation incidents.

This is one way to look at countermeasures and mitigations. It’s a useful way to examine the space of possible actions, in the same way that a naval officer learns about ‘standard’ manoeuvres like the Crazy Ivan, and how to think about detecting and mitigating for them.

Disinformation, like war, isn't a linear process: that there are techniques in play that work and are likely to be used is just the first level of understanding what could and might be done. Good incident creators are also artists (yes, yes, there's a reason it's called "the Art of War"), understanding the basic techniques and constraints, and knowing how to adapt them into a flow of actions that becomes difficult to counter with a simple rulebook. These masters still need to know the basics though.

4.1.5 Workshopping Counters

Day 1

- Introduce what MisinfoSec_WG has done, why we've done it, and what we have to show. Introduce AMITT; review stages and techniques
- Workshop/hands on "Blue Team" to build the responses part of the framework
- Create 5-7 five-person multi-disciplinary teams each responsible for creating a collection of counters for up to 10 of the 54 identified techniques

Day 2

- Introduce ISAO concepts and how they connect to misinformation
- Workshop/hands on design of ISAO network support
- Workshop/hands on exercise testing responses and network concept together
- Wash-up and next steps

4.2 DISARM Countermeasure components

When organising countermeasures, there are a few questions to ask:

- What does this counter do? Is this a mitigation, and what does it do: does it stop a technique being effective, moderate its effect or do something like delay its effect whilst other measures are put in place?
- Who can do this? What skills and resources do they need to have a chance at success? What risks do they take in doing it and how can those be both explained and minimised?
- Has this been tried before? What happened that time? Are there side effects (both good and bad) to watch out for?
- Has this been used in combination with other counters? Could it be?
- What level is this counter at? Is it strategic, tactical or immediate?

Answering these questions meant adding appropriate labels and examples to each countermeasure. This subsection covers some of those labels.

4.2.1 Countermeasure types

The list of countermeasure types is a cut-down version of the US Military's Joint Publication 3-13, aka JP3-13 This descriptions of the list items appears on page I-9:

"Objectives:

> *Commanders use IO capabilities in both offensive and defensive operations simultaneously to accomplish the mission, increase their force effectiveness, and protect their organizations and systems. Fully integrating IO capabilities for offensive and defensive operations requires planners to treat IO as a single function. Commanders can use IO capabilities to accomplish the following:*
>
> 1. *Destroy. To damage a system or entity so badly that it cannot perform any function or be restored to a usable condition without being entirely rebuilt.*
> 2. *Disrupt. To break or interrupt the flow of information.*
> 3. *Degrade. To reduce the effectiveness or efficiency of adversary C2 or communications systems, and information collection efforts or means. IO can also degrade the morale of a unit, reduce the target's worth or value, or reduce the quality of adversary decisions and actions.*
> 4. *Deny. To prevent the adversary from accessing and using critical information, systems, and services.*
> 5. *Deceive. To cause a person to believe what is not true. MILDEC seeks to mislead adversary decision makers by manipulating their perception of reality.*
> 6. *Exploit. To gain access to adversary C2 systems to collect information or to plant false or misleading information.*
> 7. *Influence. To cause others to behave in a manner favorable to US forces.*
> 8. *Protect. To take action to guard against espionage or capture of sensitive equipment and information.*
> 9. *Detect. To discover or discern the existence, presence, or fact of an intrusion into information systems.*
> 10. *Restore. To bring information and information systems back to their original state.*
> 11. *Respond. To react quickly to an adversary's or others' IO attack or intrusion All IO capabilities may be employed in both offensive and defensive operations."*

Action types *exploit, influence, protect, restore and respond* weren't viewed as immediately relevant to disinformation work.

4.2.2 Response Actors

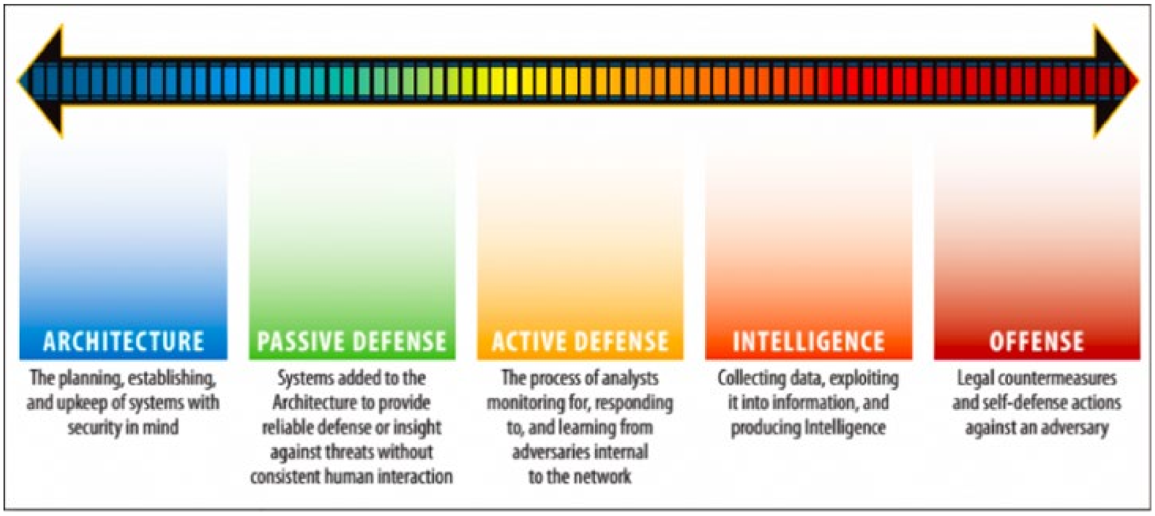
Describing actions is great, but actions only work if someone does them. There are many entities in the space of being affected by and analysing disinformation campaigns; not so many entities in the space of being able to, willing to, legally allowed to, or actively responding to disinformation. Entities who could respond include social media platforms, other organisations, civil society, media organisations, governments, militaries and individuals. There are also other stakeholders who could be persuaded or find it in their best interests to help reduce the prevalence of disinformation campaigns across societies.

Social media platforms have control over their own software, and usually have control over the data moving through it, and the data available on and archived in it. They also have control over who can access that software and data - or rather, over which accounts can access it. Very few social media companies are owned by individuals now

- - they tend to be accountable to business stakeholders whose motivation is, generally, profit. This means that removing disinformation from systems is often in competition or conflict with other business priorities, or may require system adaptations or rebuilds that are too costly to justify against an uncosted, unquantified, unknown damage to society.
- Other online organizations include organizations like web hosts and DNS registrars, who could help with the removal of disinformation campaign websites.
- Civil society is that connector between the people trying to help counter disinformation campaigns and the people who are subjected to them. This is where people-centre approaches like education and reporting routes for microtargeted messages and advertisements are tried.
- Media has its own disinformation problems, despite its emphasis to itself on trying to find truth. Falling media budgets, longer/faster news cycles and wide access to information about breaking stories has left individual net journalists struggling to keep up and wade through streams of information, malinformation and disinformation around events. The counters here are two-way - both journalists helping counter disinformation with new practices (e.g. “rumour” pages during natural disasters), and in better training on content ingestion and dissemination practices.
- Governments can help primarily with the regulations that companies can use to justify moving disinformation measures above other line items in their business plans. The shadier parts of government can also help with more direct action tracking down and dissuading campaign creators and amplifiers.

4.2.3 Meta Techniques

There are legal restrictions in many countries on the types of counter response that different actors can perform: for example, in the United States, the Posse Comitatus Act limits offensive actions of US military on US territory, making the lists of potential actors fraught with questions like “yes, this group of responders could use this countermeasure, but is it legal and/or moral of them to do so?”. Circumventing Posse Comitatus by using the National Guard notwithstanding, one of the first actions in answering that across multiple countries is to label counter TTPs by whether they’re offensive or not.



ARCHITECTURE
The planning, establishing, and upkeep of systems with security in mind

PASSIVE DEFENSE
Systems added to the Architecture to provide reliable defense or insight against threats without consistent human interaction

ACTIVE DEFENSE
The process of analysts monitoring for, responding to, and learning from adversaries internal to the network

INTELLIGENCE
Collecting data, exploiting it into information, and producing Intelligence

OFFENSE
Legal countermeasures and self-defense actions against an adversary

SANS scale

Information security has a framework for this too: the SANS scale, as shown above. In many cases, this was too coarse grained a scale to help with determining who could potentially use a measure, so we also tagged counter TTPs with the rough type of action they were suggesting, as seen below.

Metatechnique	Description	SANS
metatechnique	Not direct counters, but fit the SANS architectural level of countering	architecture
cleaning	Clean unneeded resources (accounts etc) from the underlying system so they can't be used in disinformation	passive
data pollution	Add artefacts to the underlying system that deliberately confound disinformation monitoring	passive
daylight	Make disinformation objects, mechanisms, messaging etc visible	passive
diversion	Create alternative channels, messages etc in disinformation-prone systems	passive
resilience	Increase the resilience to disinformation of the end subjects or other parts of the underlying system	passive
scoring	Use a rating system	passive
counter messaging	Create and distribute alternative messages to disinformation	active
dilution	Dilute disinformation artefacts and messaging with other content (kittens!)	active

friction	Slow down transmission or uptake of disinformation objects, messaging etc	active
reduce resources	Reduce the resources available to disinformation creators	active
removal	Remove disinformation objects from the system	active
verification	Verify objects, content, connections etc. Includes fact-checking	active
targeting	Target attackers or components of a disinformation campaign	offense

DISARM Blue Meta Technique categories

This provides a bridge between the disinformation types and the SANS scale.

4.3 Building DISARM-based Playbooks

A collection of countermeasures is nice, but it's not going to help someone who's facing an immediate active campaign or incident. They're going to need some form of "hey, this is happening, here are things you could try and what might happen" guides.

One of the things that reading through the counters spreadsheet surfaces is the sense of who is doing what to whom with which resources? For example - we have a lot of entries that look something like "tell x about y". Which is great, but that assumes that y can do something about
x. After a while this starts to look like pieces of a stix graph itself - we have actors (or types of actor), artefacts and techniques in play, connecting to and relying on each other. Content takedowns, for instance: these can only happen if the people capable of doing the takedowns know about the content, and the people who know about the content tell them about it. We may also have a componentwise, piece-together set of responses to be built. To start with, mapping out who is doing what to whom with which resources, and which assumptions about actions and outcomes might go a long way in reducing our 200+ listed counters down to a manageable tactical set.

4.4 Further Reading

Must-reads on counters

- [Stray19] Jonathan Stray, "Institutional Counter-disinformation Strategies in a Networked Democracy", WWW 2019 (video)
- The war on pineapple: https://www.dhs.gov/sites/default/files/publications/19_0717_cisa_the-war-on-pineapple- understanding-foreign-interference-in-5-steps_0.pdf
- Chapter 7 of https://www.foreign.senate.gov/imo/media/doc/FinalRR.pdf

General references on counters

- https://ukraineelects.org/live-updates/page/4/
- https://navigator.oii.ox.ac.uk/resources/?resource_filter%5Bsubject%5D%5B%5D=disinf ormation-counter-strategies#
- https://www.climatechangecommunication.org/wp-content/uploads/2020/10/DebunkingHandbook2020.pdf

5 Multi-Player Game Models: design and philosophy

One potentially fruitful model of disinformation is as a game where multiple players on both red and blue teams compete and cooperate for resources, using the TTPs from the DISARM framework and DISARM counters models. In 2020, CogSecCollab ran weekly red team exercises, usually based on incidents the team was tracking or countering online. These exercises used the AMITT TTPs, meta-techniques and STIX objects, with realistic estimates of red and blue team resources, to anticipate new disinformation narratives and moves. These were used to help prepare mitigations and watches for future incidents, and draft a “Doctrine for countering disinformation”.

Much of this work was on the operational level, using the DIMEFIL model for geopolitics and business, and TTPs to model manoeuvres in those spaces.

Critical resources included:

- Resources
    - Transmission media
    - Audience
    - Message generation (narratives?)
    - Manhours
    - Intelligence, Access, Capability
    - Credibility
    - “Money, money, and money” (Trivulzio)
- Message
- Credibility
- Access/Audience
- Temporal (timelines, deadlines)

This work used a single Centre of Gravity. The most critical resource we found was time, e.g. to delay, scope, front-run, etc.

Other recent work on DISARM and multi-player games plots the disinformation red team and blue team TTPs for an incident together, and tracks their connections and potential effects on each other.

5.1 Further Reading

Susan Young, Dave Aitel, “The Hackers Handbook”

6 DISARM Trials and Implementations

In 2020, we used AMITT in live and test disinformation defence deployments.

6.1 AMITT MISP Implementations

AMITT was implemented, tested and used in two MISP instances:

- The CogSecCollab MISP instance, used for testing by both CogSecCollab and other groups trialling the AMITT framework.
- The Covid19 MISP instance, used by groups around the world to share threat intelligence about Covid19 information security issues.

The CTI League's Disinformation team, led by CogSecCollab team members, worked with the Covid19 instance, adapting tools, processes and models to fit a team handling large volumes of information at rapid speed. Innovations added for the CTI League included

- A full set of social media objects
- A one-line command to push information about a social media artifact up to MISP.

6.2 Related Work

CogSecCollab leads also helped start and chair the DEFCON AI Village (a village dedicated to work on the interface between information security, machine learning and artificial intelligence). One of the pieces of work aided by CogSecCollab was the 2019 Rootzbook misinformation challenge, designed as a simulation exercise to help young hackers understand the processes behind disinformation and botnets.

7 Further Work

There are many things to add to the DISARM models. These include:

- Commentary: How to measure effectiveness. The importance of information sharing for detecting campaigns early. More about the DE of ABCDE and how it links to DISARM risk management models.
- Analysis: The use of natural language processing, social graph analysis, and propagation patterns on raw data.
- Counters: Difficulty of counteracting entrenched beliefs directly, information aikido, disrupting the coordination of meme/conspiracy attacks. Adding anti-harassment and counterterrorism models. Lots of approaches that are only pieces of the puzzle, or intractable/unscaleable.
- Potential for AI/ML approaches to detection and automated countermeasures.

There are also activities that can help ratify, correct, and suggest further work on the models. The most important of these is to convene users and designers to work through the proposed changes to DISARM.