Chapter 15 Taking Action

Introduction

The point of real-time disinformation tracking is to be able to do something about it. Our basic actions include:

- Not acting
- Direct action.
- Asking someone directly connected to us to take action
- Reporting to someone not directly connected to us, so they can investigate and decide whether to take action.

Not acting. This is always an option: we should always ask if we should act, and if we want to act \(and if not, what are the ethical ways we have to discharge responsibilities like having the datasets that we have\).

Direct action: there are many small things that a team could do to disrupt a disinformation incident. These include:

- Flooding a disinformation hashtag or group with alternative information \(be careful with this because if the original intent was confusion, you might be adding to it\) etc

Asking someone connected to us to take action

- Reporting a suspicious domain to registrars. If we do this, it's on us to gather information to help them - e.g. screenshots of selling bleach 'cures' etc etc
- Reporting to law enforcement. Escalating to law enforcement is appropriate especially if there is risk of physical harm, but use this route wisely.

Reporting to someone not directly connected

- This is most likely with the large social media platforms. We're going to find bots and botnets; we won't be able to remove them ourselves, we will be able to report them to platforms. It'll help if we have that reporting mechanism set up ahead of time.
- Takedown requests will need a reason, the easiest of which is violations of platform terms of service. This is about pointing platforms at incidents, artifacts and behaviors they might not have detected already; it's also about countering disinformation incidents: we are not censors, and should always view data in terms of "what is happening" rather than "I disagree with this post".

Action

What we want to do with an incident is disrupt it as much as possible. If we can stop it completely, that's a big win, but generally, we're after disruption. CogSecCollab has a long-list (here: https://github.com/cogsec-collaborative/amitt/blob/master/tactic_counts.md) of the things we can do to disrupt incidents at different stages of the disinformation killchain (https://github.com/cogsec-collaborative/amitt_framework - that, and DFRlab's object labels https://github.com/DFRLab/Dichotomies-of-Disinformation are what we're using in the MISP reporting), but frankly it's still messy so at this stage it's better to put our hacker hats on and think "which artefacts (observable objects) do we have in this incident, and what can we do to make them less effective?"

Examples: are there URLs pushing out covid5g disinfo? Are there social media accounts and groups pushing out covid5g disinfo? If we gather evidence on these, we can get that to the social media companies. Are there botnets involved \(yes, yes, I said the b word, but they're part of this too\)? Can report those too. Etc etc \(and I suspect many of you have etcs CogSecCollab didn't think of when they created that counters repo\).

This is the practical part of incident handling. We track an incident until the underlying incident stops or slows significantly \(or the event it's building up to has passed\), or until we've done as much as we believe we can to counter it, or know that there are other teams dealing with it.

Disinformation counters are much more than "remove the botnets" and "educate people". For most incidents, there are a variety of things that can be done about the incident, its creators, the objects used in it, and the tactics and techniques used. We've collected a few \(well, a couple of hundred\) suggestions for technique-level counters at https://github.com/cogsec-collaborative/amitt - we're expecting to uncover a bunch more as more infosec people do disinformation.

Countermeasures

"Countermeasures are that form of military science that, by the employment of devices and/or techniques, is designed to impair the operational effectiveness of enemy activity. Countermeasures can be active or passive and can be deployed preemptively or reactively." - JP 3-13.1 , Information Operations - Joint Chiefs of Staff

The MisinfosecWG collected and designed countermeasures against AMITT tactics and techniques in 2019. We did four main things to get our list of 200-odd counters.

- Existing: We looked for existing countermeasures, in incidents, literature and examples.
- Tactic-based and Technique-based: We ran a workshop, where we used the AMITT framework to create counters to both the tactic stages, and to specific techniques within those stages. This centred around a courses of action matrix; basically we gridded out response types and tactic stages, and asked people to post ideas into each grid square. \(our team was known for using all the postits in the building\).
- Doctrine-based: And we looked at influence operations as resource-limited games, and described the counters that we could use to deplete or exhaust disinformation resources.

We found quite a few existing counters, beyond the obvious "take down the botnets" and "educate people" ones. Examples included:

- The Macron election team's email honeypots,
- US Cybercommand blocking the Internet Research Agency's internet access during the 2018 midterm elections.

Effects-based countermeasures

Our courses of action matrix used a subset of the effects listed in JP3.0:

- Detect: find them. Discover or discern the existence, presence, or fact of an intrusion into information systems.
- Deny: stop them getting in. Prevent the adversary from accessing and using critical information, systems, and services.
- Disrupt: interrupt them. Break or interrupt the flow of information.
- Degrade: slow them down. Reduce the effectiveness or efficiency of adversary command and control or communications systems, and information collection efforts or means.
- Deceive: divert them. Cause a person to believe what is not true. military deception seeks to mislead adversary decision makers by manipulating their perception of reality.
- Destroy: damage them. Damage a system or entity so badly that it cannot perform any function or be restored to a usable condition without being entirely rebuilt.
- Deter: discourage them.

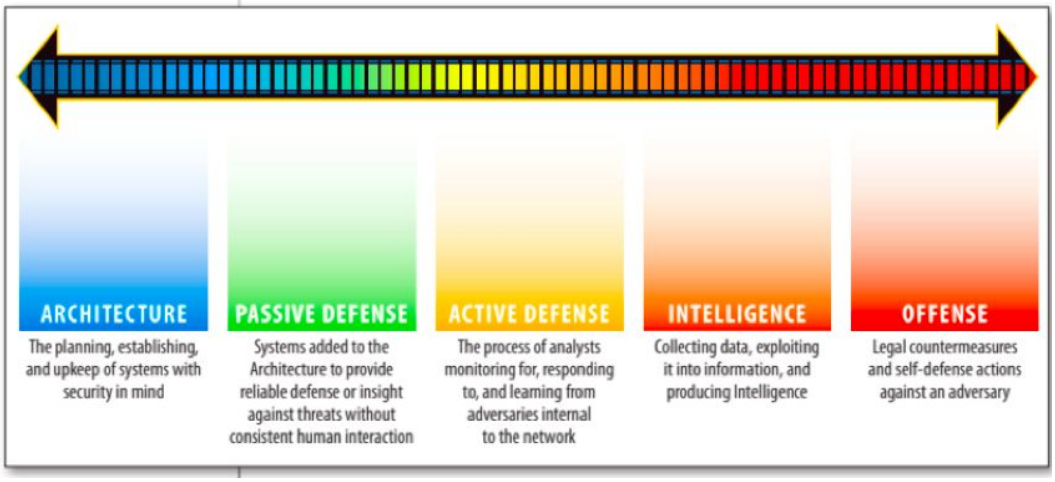
We included Detect because that's what everyone was doing - looking, not reacting, and we wanted them to get that out of their systems. We added Deter to the list as a potentially useful category too.

AMITT metatechnique courses of action

	ALL	D1 detect	D2 - Deny	D2 Deny	D2 deny	D3 Disrupt	D3 disrupt	D4 Degrade	D5 Deceive	D6 Destroy
cleaning	0	0	0	2	0	1	0	1	0	0
countermessaging	0	0	0	3	0	7	1	4	0	1
data pollution	0	0	0	0	0	1	0	4	1	0
daylight	0	1	0	5	1	8	0	2	0	1
dilution	0	0	0	0	0	5	0	1	0	0
diversion	0	0	0	2	0	10	0	2	3	0
friction	0	0	1	12	0	6	1	6	0	0
metatechnique	4	0	0	3	0	6	0	0	0	0
reduce resources	0	0	0	2	0	1	0	1	0	0
removal	0	0	0	14	1	4	0	0	0	0
resilience	0	0	0	8	2	7	0	7	0	0
scoring	0	0	0	7	0	0	0	0	0	0
targeting	0	0	0	1	0	6	0	3	0	0
verification	0	0	0	2	0	1	0	0	0	0
TOTALS	4	1	1	61	4	63	2	31	4	2

AMITT countermeasures, listed by metatechique

The result was a set of mitigations and countermeasures that we labelled by the tactic, technique, agents who could carry them out, and a metacategory that we used to sort through looking for duplicates.



SANS scale

We also thought about these effects in terms of the SANS scale for responses - e.g. whether they were architectural changes to the underlying ecosystem, defence, intelligence gathering or offence.

The AM!TT effects-based countermeasures work can be found in https://github.com/cogsec-collaborative/amitt. We're still getting the countermeasures into usable order: you can track our progress in this repository, through the grids that we're using to think about the types of blue-team actions that can be \(and are\) used against disinformation.

Example: CTI League

The CTI League uses effects-based counters: reporting to law enforcement, platforms, and registrars, with the CogSecCollab helping to set up the RealityTeam counternarratives group to help counter rapidly-evolving narratives.

Doctrine-based Countermeasures

"A disinformation campaign is made up of resources and infrastructure and operates over time, with time as a universal scarcity." - The Grugq

We need some way to make counters work well together. We have a set of counters but we need a way to understand how and when to use them. SJ, Grugq, Pablo started a conversation about resources, infrastructure, and that time is scarce. Operations are rooted in the real world

and the critical elements required in the real world limit these actors. Actors need money, people, organization, knowledge and capabilities and the time to make things happen.

When we talk about a counter for a technique we just don't want any viable counter. We want the counter that's most appropriate to our achieve our objectives within the current environment. Maybe that's to destroy some capability, or cost the adversary money, waste their time in some way that is impactful to that adversary based on that actor's state in the real-world.

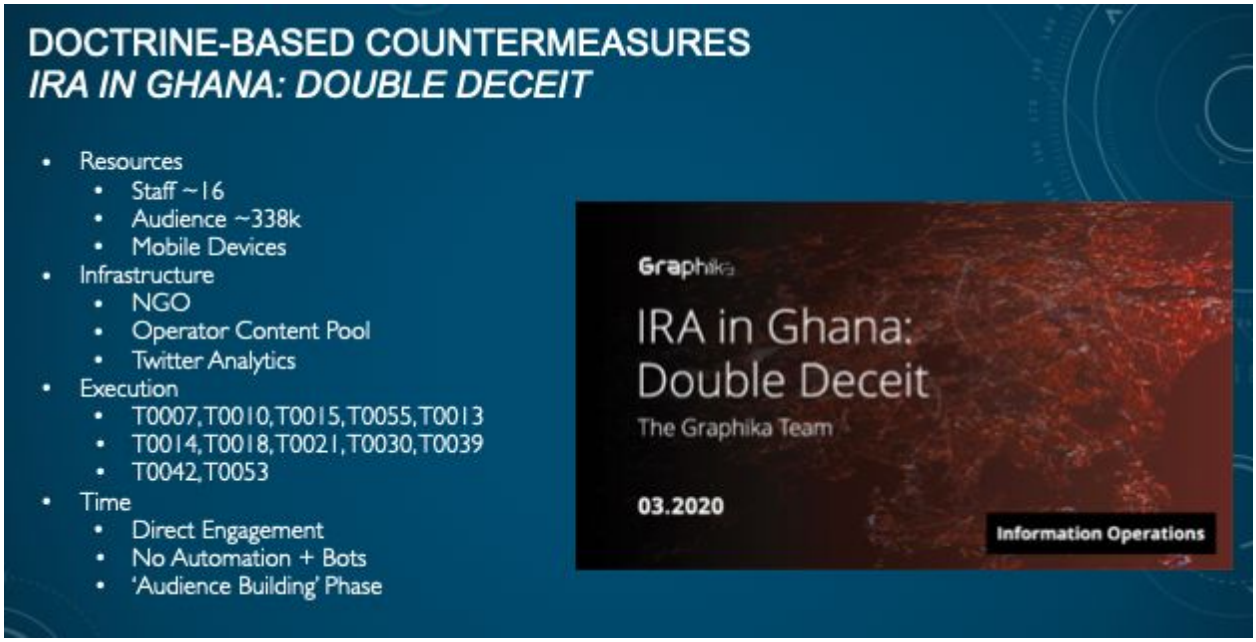
So what are the critical elements we want to affect? We like to get it RITE:

- Resources: material things, money, messages, audience
- Infrastructure: media/medium, administration, observation
- Time: speed, capacity, "mythical person month"
- Execution: actors, capabilities, strategies

Now we can build a course of action matrix based on critical elements and combine it with the course of action matrix for Amitt techniques. When action is taken on critical elements, a capability's capacity to influence a target audience is affected in some strategically significant way. Effects on disinformation capabilities can now be grouped into three main classes;

- those which exhaust the capability's resource dependencies;
- those which decrease a capability cost-effectiveness; and
- those which exhaust the adversary's capacity to use a capability in a timely manner.

Example



Double Deceit incident components

That's the theory, can we apply that in practice? Turns out we can. Double Deceit is interesting to a critical elements based doctrine of counters for several reasons: Small operation: 16 people around a table; essentially a bunch of kids with phones. complicated situation for a defender - can't target an NGO without understanding who they are and what they're doing; they might be legitimate, but appear adversarial in nature.

So what can we do the next time we detect this pattern? Banning the accounts and alerting them just teaches them about our own capabilities. Ideally we want to make them ineffective but allow them to operate. Target critical elements the adversary is weak in and turn this IO into a resource sink for our adversary.

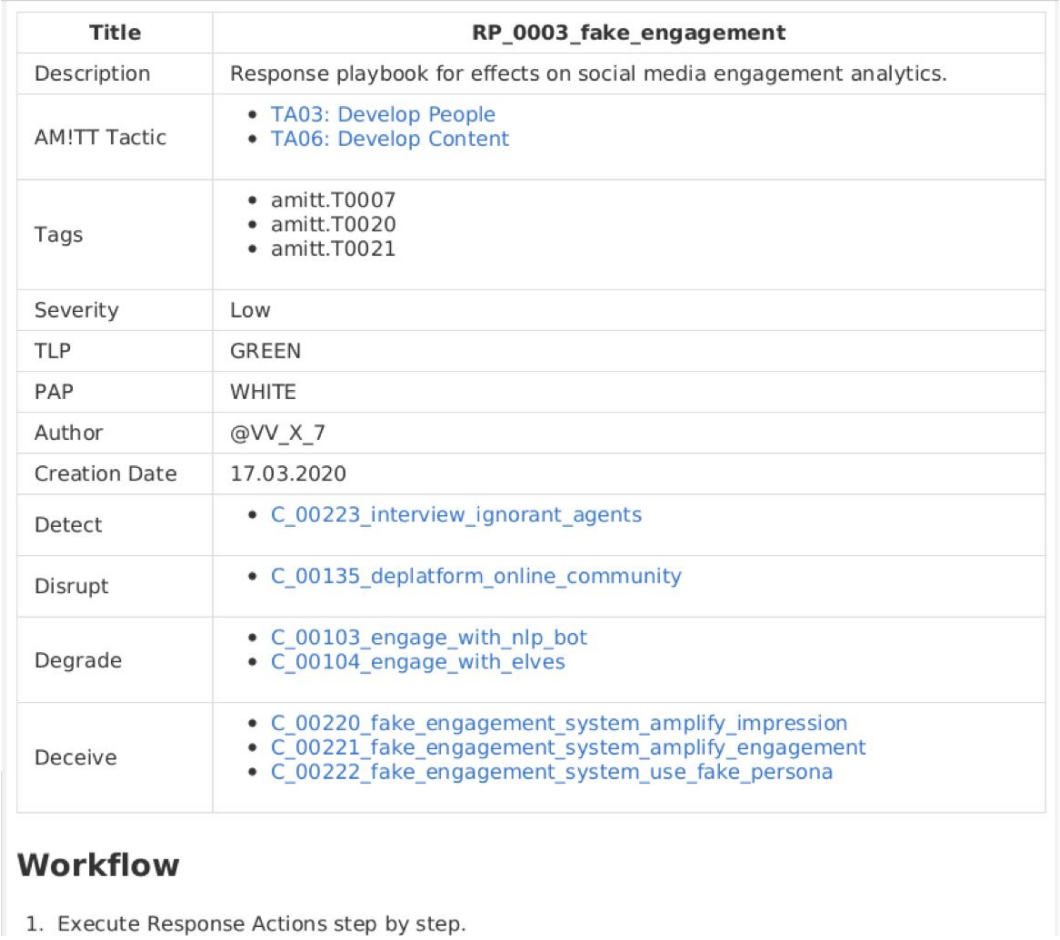
The critical elements of this operation can guide our application of counters. For example,

- this a highly time constrained operation. No bots. Direct engagement. Bottlenecks in how we could expect them to react to our counters. We can slow them down and make them ineffective.
- Another vulnerability is that they relied on Twitter Analytics and understanding retweets, messages. It appears that their KPIs are all contingent on social media platforms giving them accurate results.

Playbooks

Threat Intelligence playbooks work towards a goal. We have something we want to protect, achieve, deny etc. Playbooks can build complex responses to disinformation events. Can tell us how to respond to an adversary given a set on conditions and objectives.

In double deceit, we saw time was one vulnerability. Double Deceit was also vulnerable in how it collected and used analytics. If we want to disrupt that, we could do something like a fake engagement playbook - a set of actions to disrupt that.



Title	RP_0003_fake_engagement
Description	Response playbook for effects on social media engagement analytics.
AM!TT Tactic	• TA03: Develop People • TA06: Develop Content
Tags	• amitt.T0007 • amitt.T0020 • amitt.T0021
Severity	Low
TLP	GREEN
PAP	WHITE
Author	@VV_X_7
Creation Date	17.03.2020
Detect	• C_00223_interview_ignorant_agents
Disrupt	• C_00135_deplatform_online_community
Degrade	• C_00103_engage_with_nlp_bot • C_00104_engage_with_elves
Deceive	• C_00220_fake_engagement_system_amplify_impression • C_00221_fake_engagement_system_amplify_engagement • C_00222_fake_engagement_system_use_fake_persona

Workflow

1. Execute Response Actions step by step.

Example playbook \(fake engagement\)

A fake engagement playbook looks something like this. We list the relevant adversary capabilities and the set of appropriate counters to achieve our objective. What we're doing here is building a playbook that lists the possible effects we can have on the adversary capabilities and which counters we need to use to achieve that effect.

Our next step is integration of critical elements to guide decisions to deny, degrade, etc., by the type of resources required for the capability. For example, we could degrade adversary use of analytics by targeting time or resource sensitive requirements for those analytics. At which point we start getting into game theory.

Practical Countermeasures: External Actions

We can't always act ourselves, but we often know an organisation or group that can, given direction on where to look, and/or reliably gathered evidence. Groups like CS-ISAO also need ways to share disinformation information quickly between organisations.

Government

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.

Jonathan Stray, “Institutional Counter-disinformation Strategies in a Networked Democracy

Jonathan Stray's survey on government response is useful here.

Practical Countermeasures: Reporting

Reporting inside the League

If you know which organisation you need, use the /list_orgs and /list_contacts \[org\] slack command to find the person you need. More generally, look at the channels guide in the League handbook to see the right channel to report an incident or component to.

Reporting to law enforcement from the League

You can open an LE escalation ticket using the /lenew command

Reporting to platforms

Reporting to social media

- Reddit: https://www.reddit.com/r/redditsecurity/
- Twitter: report-twitter-impersonation and twitter-rules
- Facebook: How to Report Things on Facebook
- Linkedin: Reporting Inaccurate Information on Another Member's Profile
- Instagram: https://help.instagram.com/1735798276553028
- YouTube: https://support.google.com/youtube/answer/2802027
- Google: is going to take some digging Avoid and report Google scams - Google Help

Pinterest

- Fast: https://help.pinterest.com/en/article/report-something-on-pinterest
- Slower: report on https://help.pinterest.com/en/contact?page=about_you_page - you'll need a Pinterest account to do this from.
    - Choice is porn, violence, hate speech, self harm, harassment/ exposed private information, spam; currently going with either hate speech, violence or harassment as appropriate.
    - Has an image filesize limit of 2MB
- community guidelines are https://policy.pinterest.com/en-gb/community-guidelines

Reporting websites

If you've found a website or ring of websites, teams you can report it to \(with supporting notes\) include registrars and the lists used by adtech and other sites to check the types of sites that they're passing money through.

Site lists:

- Global disinformation index
- Media bias fact check
- Unreliable News repo , Cred Score (hypothesis), Fact-check Feed (articles by US fact-checkers, 2016–present), Fact Checkers tool, News Netrics media site performance metrics.
- https://iffy.news/fact-check-search/

Practical Countermeasures: Direct Action

Some groups \(e.g. the kPop stans\) have taken direct actions against disinformation, including flooding hashtags with external material \(band photos etc\).

When you act in a disinformation space, you're acting in an environment, with a lot of other humans and machines in. And what you can end up in is a multiplayer game, where you're each acting in response to each other, and playing off against each others' resources. Be aware of this if you choose this route.

Further Reading

- Training end-users about disinformation
    - https://getbadnews.com/#intro - game to train people on how disinformation works
    - CrashCourse media literacy videos