Companion Guide to the 2019 ‘Blue’ workshop output

Objective

The goal of this document is to offer some guidance to anyone using the ‘Blue’ framework output of possible countermeasures from a workshop of volunteers in 2019. Given its links to the Red Framework we recognize that parts of this ‘Blue’ thinking may be useful to some of our users and so we continue to make it available. However, parts of this output are highly problematic from the point of view of democratic values and ethical principles, if read without context and qualification. For this reason, the DISARM Foundation does not endorse this ‘Blue’ framework, as it stands, and advises caution to anyone using it. We have written this document for those users, to provide some considerations for its use, based on democratic values and ethical principles.

Intended Audience and Scope

This document aims to provide ethical considerations for anyone involved in protecting the information environment from manipulation and online harm. The intended audience is broad. However, this document limits its analysis of the acceptability of responses to information manipulation to non-governmental actors. Value judgments are made about the level of acceptability of actions that non-governmental actors might take based on international human rights principles. Value judgments about actions that governments might take is beyond the scope of this document[1].

Disclaimer

The information contained in this document is provided for informational purposes only and should not be construed as legal advice on any subject matter. You should not act or refrain from acting on the basis of any content included in this document without seeking legal or other professional advice. The contents of this document contain general information and may not reflect current legal developments in your jurisdiction or address your situation. The DISARM Foundation disclaims all liability for actions you take or fail to take based on any content in this document.

Background

The beginnings of the DISARM Framework were in 2018, when a group of like-minded individuals decided to help ‘frame’ the problem of disinformation. They saw that those creating disinformation campaigns had only a limited need for coordination. On the other hand, many different groups can be affected by their campaign. And if those affected do not share a common understanding of what is happening, they will always be at a disadvantage. Hence the need for a ‘framework’, to help provide that common understanding.

In 2019, a wider group of volunteers participated in workshops to build the framework, titled AMITT: Adversarial Misinformation Influence Tactics & Techniques. This mirrored the approach of an existing cybersecurity framework, which allowed users to understand a) what is happening (this was summarized in a ‘Red’ framework) and b) what could be done about it (to be summarized in a ‘Blue’ framework). The AMITT Red framework was published to be available and free to all, and a wide variety of organizations began using it to categorize the disinformation campaigns they were seeing.

The DISARM Foundation launched early in 2022 after the framework was renamed from AMITT to DISARM — Disinformation Analysis & Risk Management Framework. The new entity, with its very limited resources, has been focused on the Red framework: continuing to improve its usefulness, and ensuring it is updated and kept open and free to use by the community of users. We have not had the resources to build on the output from the 2019 workshop that focused on creating the 'Blue' framework of possible responses.

The 2019 'Blue' workshop aimed to ensure as much academic completeness as possible by capturing all observed and potential courses of action for defending against disinformation. These included courses of action seen in countries with different ethical values (e.g. the 'Blue' workshop output included "censorship" because that's what China was doing at the time). By taking this approach the workshop organizers also aimed to make it easier for democratic governments and societies to make informed decisions about where the boundaries should be and what limitations should be placed on 'Blue' actors. The 'Blue' workshop output was thus a values-agnostic framework designed to track all possible defender actions. It was not intended as a set of recommendations. Due to the resource constraints of the volunteers participating in the workshop, the 'Blue' workshop output did not provide cautionary guidance for users other than simply labeling some courses of action 'not recommended'.

This document now provides further guidance and context to the 2019 'Blue' workshop output. Going forward, the DISARM Foundation in 2024 is creating the vision for a new 'Blue' framework – based on further workshops we held in 2023, and also based on democratic values and ethical principles – which will, when complete, be placed alongside the existing Red Framework, so the two can work hand-in-hand. When complete, we will have fully realized the original vision of 2018.

How to Use This Guide

The DISARM Foundation does not endorse the use of the 2019 'Blue' workshop output for anything other than the purpose for which it was intended i.e. to describe and track what actors do to defend against disinformation campaigns directed against them. We recognize, however, that some of our users find the 'Blue' workshop output useful as a potential checklist of possible actions that they themselves might take. To those users we advise caution when using the 'Blue' workshop output, as many actions are inappropriate and do not comply with democratic values, and so we offer this guide to highlight those actions we believe are problematic and to explain why we believe they are problematic.

Please keep the following factors in mind when using this guide:

- There is no one-size-fits-all framework for defense against disinformation. When considering defensive actions, every democratic society must apply its own distinct laws and norms based on its own unique history and culture, while complying with international treaties and customary international law.
- The boundary of what is an acceptable response to disinformation varies not just according to local laws and norms but also according to context. For example, deception is generally unacceptable in peacetime but can be an acceptable tactic in warfare. We have tried to anticipate the different contexts that users might face when considering each countermeasure and have discussed these under "Ethical and Legal Considerations", but it is simply impossible to think of every potential scenario. Users must consider their own unique context when assessing the acceptability of countermeasures.

- Classifying potential actions into different levels of ethical acceptability is inevitably a subjective exercise. We had disagreements amongst ourselves when compiling this guide. In the end we erred on the side of caution, placing the acceptability bar high in terms of demanding transparency and protecting freedom of speech. But this is only a guide. Every user must consider for themselves what actions would be ethical, relevant, proportionate, and appropriate, given the unique legal, cultural, and normative context in which they are acting.

Balancing Freedom of Expression and Freedom from Harm

We recognize that each nation, jurisdiction, and community may draw the line differently when balancing freedom of expression and freedom from harm. As a group of individuals residing in the US and Europe, we are keenly aware of the distinct approaches taken on each side of the Atlantic.

Our overall frame of reference is the United Nations Declaration of Human Rights[2], but the following analysis also leans heavily on current US First Amendment jurisprudence. This involves not just the text of the First Amendment itself, which states "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances", but also the marketplace of ideas framework proposed by revered Supreme Court justice Oliver Wendell Holmes, Jr., in his 1919 dissent in Abrams vs United States[3].

The principal idea is that democratic citizens should be free to express their ideas openly, and that, through the rigorous debate of speech and counter-speech, the truth will prevail[4]. As decided by the US Supreme Court in 2012 in United States v. Alvarez, lies are an inevitable by-product of such a marketplace of ideas, because prohibitions on lies would have a chilling effect on free speech[5]. In the US system, therefore, lies are constitutionally protected, except in cases involving "defamation, fraud, or some other legally cognizable harm associated with a false statement, such as an invasion of privacy or the costs of vexatious litigation"[6]. Furthermore, in defamation lawsuits in the US the onus is on the plaintiffs to prove falsity or even, in some cases, malice, because, if it were easy to bring a defamation lawsuit, this could also have a chilling effect on free speech[7].

Committing to Democratic Values

In December 2023, we reached out to InfoEpi Lab for advice and guidance on establishing a values-driven and ethical approach to countering disinformation. Following these discussions, we propose that our users commit to the following democratic values when defending against disinformation, influence operations, or online harm.

Transparency and Accountability: Any actions taken should be transparent and accountable to the public. This includes disclosing the sources and intentions behind information campaigns and mechanisms for public oversight and critique. Organizations working to counter disinformation should not have hidden relationships that they would not stand by publicly.

Nonmaleficence: The guiding principle of counter-disinformation activities is "Do no harm." This principle ensures that actions taken are not just effective but morally sound and respectful of the rights and dignity of individuals.

Upholding Free Speech and Thought: Counter-disinformation efforts must prioritize the protection of free speech and free thought. This means avoiding tactics that suppress, censor, or manipulate public discourse. Instead, the focus should be on enabling informed decision-making by providing accurate, clear, and accessible information.

> *Everyone has the right to freedom of thought, conscience, and religion; this right includes freedom to change his religion or belief and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship, and observance.*
>
> *Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive, and impart information and ideas through any media and regardless of frontiers.*
>
> – United Nations. (1948). Universal Declaration of Human Rights.

Although no society has absolute free speech, censorship should be recognized as inherently undesirable. To understand where the norms around freedom of speech exist in a given society, examine the laws surrounding it. For example, child sex abuse material is not protected speech, and publishing it is a criminal act. One can be sued for defamation or libel. Parties responding to or engaging with society should know and observe speech norms in a given society.

Respecting Privacy and Autonomy: Any data collection must respect individual privacy and autonomy. Data should come from legal and publicly available sources. Any published data should be appropriately anonymized.

Rejecting Unethical Influence

Borrowing again from InfoEpi Lab, we propose that our users commit to refraining from defensive actions or response measures that involve unethical influence. “Unethical Influence” refers to a broad range of manipulative tactics and practices aimed at altering an individual’s perceptions, beliefs, or behaviors through morally questionable or outright deceptive means. Unethical influence is often characterized by a lack of transparency, a lack of respect for autonomy, or the lack of a fair presentation of information. It undermines the principles of informed consent and free will, leading to decisions or beliefs that might not reflect the individual’s valid preferences or best interests.

This concept encompasses various methods, including but not limited to:

Deception: Utilizing false information, misleading statements, or presenting facts out of context to sway someone’s understanding or decision-making in a way that benefits the influencer at the expense of the influenced.

Exploitation of Biases and Mental Heuristics: Taking advantage of inherent cognitive biases or mental shortcuts that people use to process information. This could involve playing on common tendencies like confirmation bias (favoring information that confirms existing beliefs) or the bandwagon effect (conforming to what others do).

Misleading Communication: Deliberately crafting ambiguous messages containing half-truths or framing them in a way that leads to misinterpretation or a skewed understanding of the situation. Another common example is misleadingly presenting a messenger, such as interviewing a medical doctor in an area outside their expertise to contradict relevant experts. This unethical influence tactic, pioneered by Big Tobacco, exploits the trust that many people have in medical professionals.

Use of Traditional Censorship: Imposing restrictions on free speech or access to information, typically through authoritative or institutional means, to prevent certain viewpoints or information from being disseminated or heard.

Employment of Alternative Censorship Methods: This includes tactics like targeted harassment, doxxing (publicly revealing private information), or other forms of intimidation to silence or discourage individuals from expressing their opinions or sharing information.

Psychological Manipulation: Engaging in tactics that affect emotions, fears, or psychological vulnerabilities. This could involve gas-lighting (making someone question their reality), exerting undue pressure, or using fear-mongering tactics.

Abuse of Power or Authority: Leveraging a position of power or authority to influence someone's decisions or beliefs in a way that may not be in their best interest but serves the agenda of the person in power.

Selective Information Exposure: Deliberately limiting someone's access to a full range of information, thereby shaping their perception based on a curated set of data or viewpoints.

Ethical Analysis of the 'Blue' output

We analyzed the Blue workshop output, created in 2019, using the conceptual framework outlined above: balancing freedom of expression and freedom from harm while committing to democratic values and rejecting unethical influence.

Out of 140 countermeasures we categorized 53 as "largely unproblematic", 58 as "potentially problematic", and 29 as "highly problematic", from an ethical standpoint. See Figure 1. Our detailed reasoning for the categorization of each counter is explained in Table 3, Table 4, and Table 5 below. We also provide summary reasoning based upon the two ways in which counters are grouped in the Blue workshop output: "metatechnique" and "response type". Metatechniques are more civilian in nature while response types are derived from U.S. military doctrine. See Table 1 for a detailed analysis and Figure 2 for a breakdown by metatechnique. See Table 2 for a detailed analysis and Figure 3 for a breakdown by response type.

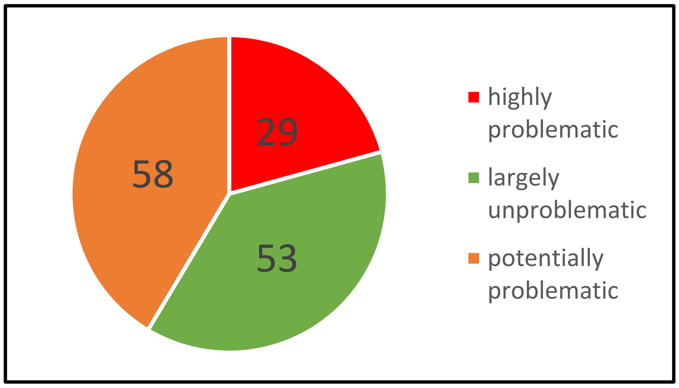

Figure 1 Ethical analysis of Blue workshop output

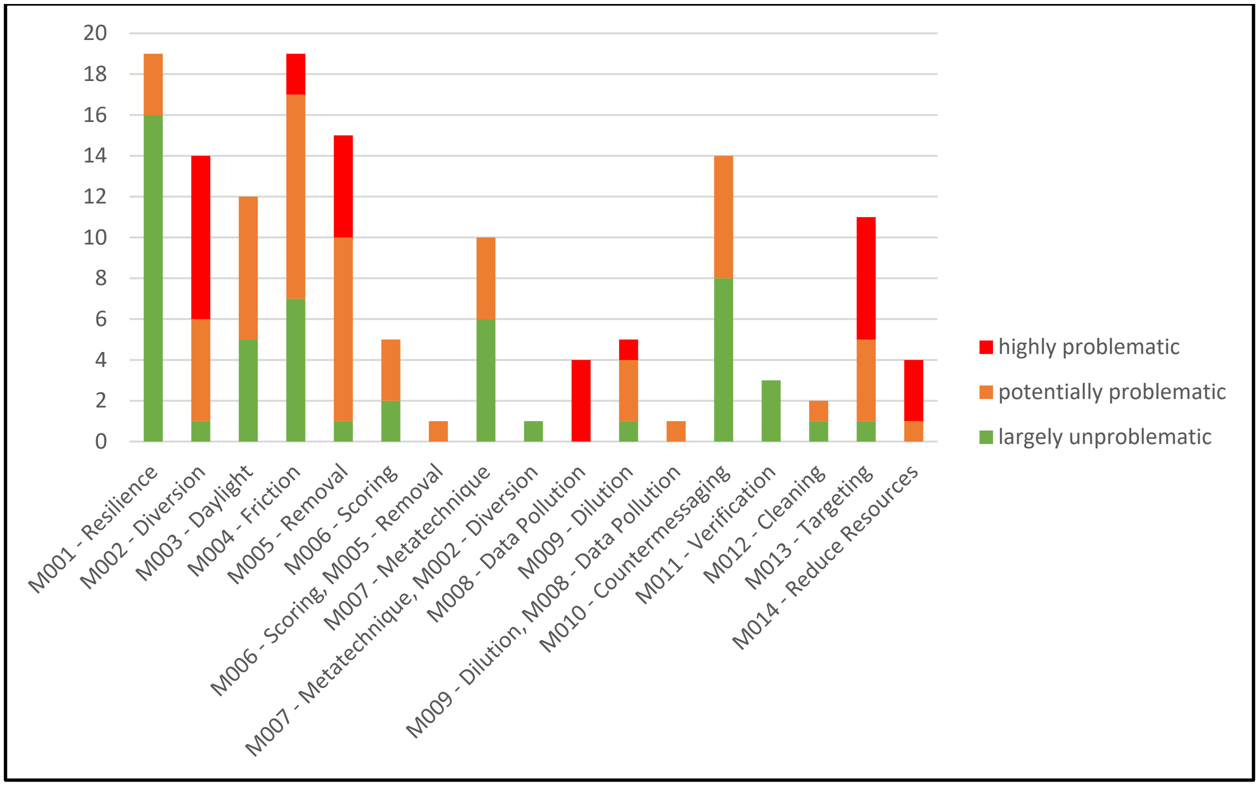

Figure 2 Ethical analysis of Blue workshop output by metatechnique

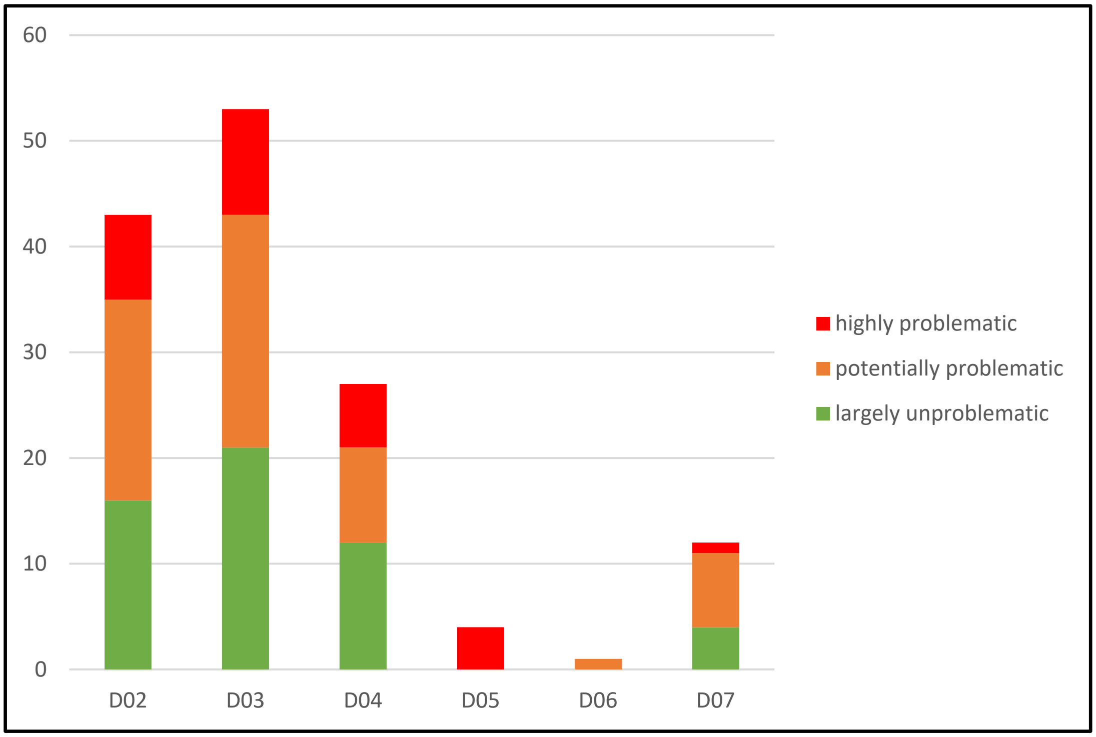

Figure 3 Ethical analysis of Blue workshop output by response type

D02=Deny, D03=Disrupt, D04=Degrade, D05=Deceive, D06=Destroy, D07=Deter

Table 1 Metatechniques

disarm_id	Name	Summary	Ethical and Legal Considerations
M001	Resilience	Increase the resilience to disinformation of the end subjects or other parts of the underlying system	Unproblematic - consists of media literacy, and ways of detecting and recovering from disinformation attacks
M002	Diversion	Create alternative channels, messages etc in disinformation-prone systems	It is highly problematic to use any kind of deception. The assumption here is that diversion would not be transparent. It is unproblematic to guide users transparently to authoritative sources of information.
M003	Daylight	Make disinformation objects, mechanisms, messaging etc visible	This is unproblematic - this is by far the preferred approach for dealing with disinformation i.e. transparency and shedding light on disinformation. It only becomes problematic where there is debate as to whether something is worthy of being labeled e.g. if it consists of false information, it may not be illegal, but it can potentially cause harm, so labeling it or shedding light on it may be advisable in certain circumstances. Labeling is adding speech, not removing it, so this is a form of counter-speech, not censorship, but some may regard this is undesirable or as impeding the free flow of information and ideas.
M004	Friction	Slow down transmission or uptake of disinformation objects, messaging etc	This is throttling or down-ranking of content, something that the platforms do with their recommendation algorithms e.g. they may downrank certain content that has been labeled by fact-checkers as false. This is potentially quite problematic since the criteria for labeling content for downranking may be highly subjective and may also not be transparent. Such labeling may be appropriate for false assertions of fact (e.g. incorrect information about where and when to vote). It is not appropriate for viewpoints or ideas. Ideally platforms allow their users to decide whether they want fact-checked content that is deemed false to be downranked. By default, Facebook downranks such content, but users can disable this downranking in their user profile. See Following Musk's lead, Youtube and Facebook are giving up on policing conspiracies - The Washington Post. The key for platforms is transparency of the terms and conditions applied by content moderators, transparency of what is moderated and how and a chance to appeal, and transparency of algorithms. Still, commercial platforms providers usually enjoy editorial discretion about what they do and do not publish and how they prioritize the content that they publish.

M005	Removal	Remove disinformation objects from the system	Potentially very problematic unless there is a clear violation of laws or terms and conditions. In general removal of accounts, takedowns of pages, channels, web domains is highly problematic unless there is incontrovertible harm that can be demonstrated. It is a matter of balancing freedom from harm with the rights to freedom of speech - when and under what circumstances do content providers lose their freedom of speech? Platforms are advised to err on the side of caution and only remove content or accounts when there is a clear violation of law or policy.
M006	Scoring	Use a rating system	Used widely in the cybersecurity industry with reputation systems, as well as blacklists and whitelists. Some of these systems are crowdsourced e.g. ratings given by shoppers to products or services, others are maintained by cybersecurity companies who specialize in malware or phishing. When applied to information in the public sphere such indices or reputation scores are more problematic, given that the definition of what constitutes harm is often less clear than in the case of cyber, in which malware and ransomware harms are more obvious. Community Notes is an example of a crowdsourced system for flagging disinformation: it works well for less polarized issues, but often fails around extreme issues. The Global Disinformation Index is an example of a proprietary ratings system which has seen backlash by conservatives in the US. Anything that is used to block content is potentially problematic, especially if the government is involved, unless it arises from content that society can largely disagree is off limits such as child pornography or gruesome terrorist propaganda involving beheadings etc.
M007	Metatechnique		This appears to involve governance actions including plans, policies, partnerships, allocation of resources, and strategic initiatives. It is potentially problematic if government involvement creates a chilling effect or an abuse of power.
M008	Data Pollution	Add artefacts to the underlying system that deliberately confound disinformation monitoring	This involves flooding the information environment with useless, spurious, or fake content, or targeting specific services with denial-of-service attacks. It is at best a disservice to the population, at worst a violation of the law.
M009	Dilution	Dilute disinformation artefacts and messaging with other content (kittens!)	At best this is counter-speech, but it may be an attempt to distract an audience from seeing certain types of content by injecting alternative but unrelated content into the environment. In that case it would create noise rather than contributing to the marketplace of ideas. Done at scale this would be closer to data pollution.

M010	Countermessaging	Create and distribute alternative messages to disinformation	This is potentially problematic unless it is done transparently. If "positive narratives" are spread in a clandestine or covert manner, then this is problematic. The key is to be transparent about what you are doing. This includes the use of automation to spread accurate and factual information e.g. about polling station locations and opening times - it is arguably ok to use automation to propagate accurate information about polling stations during an election in response to a FIMI campaign from a foreign actor who is spreading false information about polling stations.
M011	Verification	Verify objects, content, connections etc. Includes fact-checking	It is largely unproblematic to check the authenticity of both social media entities and also content - democratic societies have largely accepted fact-checking as a way of dealing with false or misleading information, even if there are debates about such systems becoming politicized. It gets more problematic if fact-checked content that is deemed false is downranked (see under Friction). Russia and other FIMI propagators are using fake fact-checking as a technique, so such systems in democracies may lose their effectiveness over time.
M012	Cleaning	Clean unneeded resources (accounts etc) from the underlying system so they can't be used in disinformation	This is also largely unproblematic and is a case of good housekeeping in terms of deactivating dormant accounts and releasing domains that are no longer paid up etc.
M013	Targeting	Target the components of a disinformation campaign	Any type of targeting of an audience or users or resources or infrastructure is problematic and should be reserved to law enforcement or intelligence agencies who have the legitimate authority to take such actions e.g. Cyber Command in the US is allowed to "hack back" against foreign cyber criminals or nation-states engaged in attacking US entities, but this capability is not legal for regular citizens to conduct
M014	Reduce Resources	Reduce the resources available to disinformation creators	Cutting off resources of disinformation creators or engaging in offensive behavior that causes them to consume their resources is highly unethical and probably illegal unless carried out by authorized law enforcement or military personnel under circumstances strictly circumscribed in law.

Table 2 Response types

disarm_id	name	summary	Ethical and Legal Considerations
D01	Detect	Discover or discern the existence, presence, or fact of an intrusion into information systems.	Detection of manipulation of the information environment should be acceptable in a democracy depending upon what is being detected, who is informed, and what they do about it. If not done transparently this can be Orwellian - Big Brother is watching you. If what is being detected is disinformation or misinformation then this needs to be clearly defined and made transparent, since such classifications can be subjective and subject to bias or even abuse.
D02	Deny	Prevent disinformation creators from accessing and using critical information, systems, and services. Deny is for an indefinite time period.	This is problematic unless the harms outweigh the right to freedom of expression according to the law.
D03	Disrupt	Completely break or interrupt the flow of information, for a fixed amount of time. (Deny, for a limited time period). Not allowing any efficacy, for a short amount of time.	This is problematic unless the harms outweigh the right to freedom of expression according to the law.
D04	Degrade	Reduce the effectiveness or efficiency of disinformation creators' command and control or communications systems, and information collection efforts or means, either indefinitely, or for a limited time period.	This is largely problematic in a democracy where the free flow of ideas includes the right to lie, given the difficulty in ascertaining truth, and the chilling effect that suppressing lies would have on freedom of expression. If degradation includes throttling or down-ranking by platforms, then this may be an acceptable application of freedom of speech by the platforms themselves.
D05	Deceive	Cause a person to believe what is not true. military deception seeks to mislead adversary decision makers by manipulating their perception of reality.	This type of action is highly problematic and unethical. As the summary suggests, such actions need to be restricted to wartime when military deception may be permitted and carried out by authorized personnel only.
D06	Destroy	Damage a system or entity so badly that it cannot perform any function or be restored to a usable condition without being entirely rebuilt. Destroy is permanent, e.g. you can rebuild a website, but it's not the same website.	This type of action is highly problematic and unethical. It needs to be limited to extreme situations such as countering terrorism or paedophilia or organized crime, in which defense, intelligence, or law enforcement are authorized by law to take such actions.

D07	Deter	Discourage.	Deterrence by denying attackers the benefits they seek or by increasing the attackers' costs may involve ethical actions (e.g. digital media literacy or sanctions against terrorists). This is a major topic of the work of the European Hybrid Center of Excellence. However, such measures can also be problematic e.g. "naming and shaming" can lead to cancel culture and the tyranny of the mob.

Table 3 Counters which are largely unproblematic

disarm _id	name	metatechnique	summary	ethical and legal considerations	tactic	response type
C00022	Innoculate. Positive campaign to promote feeling of safety	M001 - Resilience	Used to counter ability based and fear based attacks	Anticipating threats of disinformation and online harm and then inoculating target audiences is a way to educate the public and build resilience	TA01 Strategic Planning	D04
C00006	Charge for social media	M004 - Friction	Include a paid-for privacy option, e.g. pay Facebook for an option of them not collecting your personal information. There are examples of this not working, e.g. most people don't use proton mail etc.	Subscription based models can go a long way to eliminating perverse incentives	TA01 Strategic Planning	D02
C00008	Create shared fact-checking database	M006 - Scoring	Share fact-checking resources - tips, responses, countermessages, across respose groups.	Fact-checking networks now exist across democracies and have become an accepted component of the information environment. Obviously, fact-checking can be subject to the biases of the fact-checkers and can also be exploited by bad faith actors as cover for propaganda and disinformation, so full transparency and checks and balances are needed.	TA01 Strategic Planning	D04

C00009	Educate high profile influencers on best practices	M001 - Resilience	Find online influencers. Provide training in the mechanisms of disinformation, how to spot campaigns, and/or how to contribute to responses by countermessaging, boosting information sites etc.	Influencers and celebrities can serve as "useful idiots" for disinformation campaigns unless they are educated. Education needs to be non-partisan and based on democratic values and human rights.	TA02 Objective Planning	D02
C00010	Enhanced privacy regulation for social media	M004 - Friction	Implement stronger privacy standards, to reduce the ability to microtarget community members.	regulation is only one approach. Providing users of tech platforms some choice regarding the use of their private information is a first step, including the ability to opt-in. Educating users on the use of anonymity software such as privacy browsers, VPNs, TOR and to reduce their digital footprint e.g. using services such as DeleteMe is another. Regulation is needed to shed light on the clandestine collection of private information and the lack of privacy notices.	TA01 Strategic Planning	D02
C00011	Media literacy. Games to identify fake news	M001 - Resilience	Create and use games to show people the mechanics of disinformation, and how to counter them.	Games such as the Bad News Game are a wonderful educational tool for media literacy, disinformation awareness, and resilience	TA02 Objective Planning	D02
C00014	Real-time updates to fact-checking database	M006 - Scoring	Update fact-checking databases and resources in real time. Especially import for time-limited events like natural disasters.	Particularly important for crisis situations and fast-moving events in which information voids can be quickly exploited by manipulators	TA06 Develop Content	D04

C00021	Encourage in-person communication	M001 - Resilience	Encourage offline communication	Mediators involved in conflict resolution who have credibility with both parties to a conflict may be able to facilitate offline communication to help resolve a conflict. However, if this is done without the consent and trust of the parties it may come across as patronizing or elitist	TA01 Strategic Planning	D04
C00026	Shore up democracy based messages	M010 - Countermessaging	Messages about e.g. peace, freedom. And make it sexy. Includes Deploy Information and Narrative-Building in Service of Statecraft: Promote a narrative of transparency, truthfulness, liberal values, and democracy. Implement a compelling narrative via effective mechanisms of communication. Continually reassess messages, mechanisms, and audiences over time. Counteract efforts to manipulate media, undermine free markets, and suppress political freedoms via public diplomacy	This is classic public diplomacy and, as such, should be acceptable when directed at foreign audiences. When directed against domestic audiences this would need to be a non-partisan effort that accounts for freedom of speech and the fact that lying and promoting untruths has always been an integral part of democratic discourse	TA01 Strategic Planning	D04
C00027	Create culture of civility	M001 - Resilience	This is passive. Includes promoting civility as an identity that people will defend.	Civil society organizations and technology companies can play a major role encouraging a culture of civility. Examples include Boston Children's Hospital's Digital Wellness Lab (Digital-Wellness-Lab-White-Paper-Civility-Online.pdf (digitalwellnesslab.org)) and Microsoft's Digital Civility initiative (Digital Civility Index & Our Challenge | Microsoft Online Safety), including its annual report "Civility, Safety and Interaction Online" (PowerPoint Presentation (microsoft.com)).	TA01 Strategic Planning	D07

C00028	Make information provenance available	M011 - Verification	Blockchain audit log and validation with collaborative decryption to post comments. Use blockchain technology to require collaborative validation before posts or comments are submitted. This could be used to adjust upvote weight via a trust factor of people and organisations you trust, or other criteria.	Any feature that sheds light on the provenance of information is beneficial provided it is not forced upon users i.e. if it is a feature that users can enable or an unobtrusive option to get additional information then it preserves the rights of those who just want to consume the information without knowing its provenance	TA02 Objective Planning	D03
C00030	Develop a compelling counter narrative (truth based)	M002 - Diversion		This is classified as a responsetype of "Disrupt" and a metatechnique of "Diversion" but it is not an aggressive counter. It merely involves the promotion of an alternative and competing narrative in the marketplace of ideas.	TA02 Objective Planning	D03
C00042	Address truth contained in narratives	M010 - Countermessaging	Focus on and boost truths in misinformation narratives, removing misinformation from them.	Giving airtime to the truthful aspects of narratives is unproblematic	TA15 Establish Social Assets	D04
C00040	third party verification for people	M011 - Verification	counters fake experts	Exposing fake credentials or individuals pretending to be experts needs to be done thoughtfully and without acrimony or ad hominem language	TA15 - Establish Social Assets	D02
C00051	Counter social engineering training	M001 - Resilience	Includes anti-elicitation training, phishing prevention education.	Educating the public to spot and respond to attempts at social engineering or elicitation of private or sensitive information goes a long way to building societal resilience.	TA15 - Establish Social Assets	D02
C00053	Delete old accounts / Remove unused social media accounts	M012 - Cleaning	remove or remove access to (e.g. stop the ability to update) old social media accounts, to reduce the pool of accounts available for takeover, botnets etc.	This is a prudent measure that platforms could take to reduce the overall incidence of account takeover. However, such actions would typically be up to the private companies running those platforms, who may make decisions based on commercial	TA15 Establish Social Assets	D04

				factors. This counter belongs in the "best practices" category.		
C00059	Verification of project before posting fund requests	M011 - Verification	third-party verification of projects posting funding campaigns before those campaigns can be posted.	Verifying that a project is real before granting funds to the project is a prudent anti-fraud measure	TA15 Establish Social Assets	D02
C00060	Legal action against for-profit engagement factories	M013 - Targeting	Take legal action against for-profit "factories" creating misinformation.	Depending on the jurisdiction, the harms caused, and whether or not those affected have standing, this may be an effective way of raising the costs for those who attempt to cash-in on disinformation or conspiracy theories. But this also creates cost for those who are defending the information environment and initiating the legal action. Lobbying lawmakers to clarify the laws around such activities may be more effective. And lawsuits may run up against a defense involving freedom of speech, so caution should be exercised here.	TA02 Objective Planning	D03
C00062	Free open library sources worldwide	M010 - Countermessaging	Open-source libraries could be created that aid in some way for each technique. Even for Strategic Planning, some open-source frameworks such as DISARM can be created to counter the adversarial efforts.	Not clear what is meant here, but in general counter-messaging or counter-speech is exercising the rights to freedom of expression, and provided it is carried out transparently, and the source and intent is clear, this should be unproblematic.	TA15 Establish Social Assets	D04

C00073	Inoculate populations through media literacy training	M001 - Resilience	Use training to build the resilience of at-risk populations. Educate on how to handle info pollution. Push out targeted education on why it's pollution. Build cultural resistance to false content, e.g. cultural resistance to bullshit. Influence literacy training, to inoculate against “cult” recruiting. Media literacy training: leverage librarians / library for media literacy training. Inoculate at language. Strategic planning included as inoculating population has strategic value. Concepts of media literacy to a mass audience that authorities launch a public information campaign that teaches the programme will take time to develop and establish impact, recommends curriculum-based training. Covers detect, deny, and degrade.	Besides the term "degrade" this initiative has broad societal acceptance provided it takes account of the laws on freedom of speech pertinent to the population being educated. In some jurisdictions, such as the US, freedom of speech includes (with some exceptions) the right to lie and includes the right to consume lies, so education around the value of accurate information in decision-making, of the importance of verifiable data or evidence to policy-making, or the advantages of critical reasoning when evaluating information and information sources, needs to take into account these rights.	TA01 Strategic Planning	D02
C00075	normalise language	M010 - Countermessaging	normalise the language around disinformation and misinformation; give people the words for artefact and effect types.	Standardization initiatives which facilitate more efficient human communication about disinformation and misinformation or offer lexical interoperability for machines can go a long way to improving a whole-of-society approach. Examples are initiatives such as Truth in Media's "Information Quality Framework" in the US or the international efforts to standardize taxonomies within the OASIS Open Project "Defending Against Disinformation Common Data Model".	TA06 Develop Content	D02
C00081	Highlight flooding and noise, and explain motivations	M003 - Daylight	Discredit by pointing out the "noise" and informing public that "flooding" is a technique of disinformation campaigns; point out intended objective of "noise"	Shedding light on flooding and noise serves to educate the public and increases awareness and resilience	TA06 Develop Content	D03

C00112	"Prove they are not an op!"	M004 - Friction	Challenge misinformation creators to prove they're not an information operation.	While the utility of this technique is questionable the technique itself is consistent with democratic debate and the marketplace of ideas	TA08 Pump Priming	D02
C00094	Force full disclosure on corporate sponsor of research	M003 - Daylight	Accountability move: make sure research is published with its funding sources.	This follows the principle of transparency and accountability and helps readers better understand the resources, motivations, and potential biases behind the research they are consuming. This is a best practice in the counter-disinformation field followed by leading organizations such as the Atlantic Council.	TA06 Develop Content	D04
C00097	Require use of verified identities to contribute to poll or comment	M004 - Friction	Reduce poll flooding by online taking comments or poll entries from verified accounts.	In principle a mechanism to verify the authenticity of accounts to ensure the account holder "is who they say they are" should reduce the spread of fraud and misinformation. In practice, it depends on how verification is performed. The example of Twitter's (X's) "blue checkmark" is illustrative and controversial, as the verification process has changed over time and has not always lived up to expectations. At the time of writing X's Premium tier offers several added features, including higher "reply prioritization". Commercial platforms are at liberty to amend their algorithms to favor verified accounts as they see fit.	TA07 Channel Selection	D02
C00099	Strengthen verification methods	M004 - Friction	Improve content veerification methods available to groups, individuals etc.	It is great to provide users with best practices and tools for verifying the content at their disposal	TA07 Channel Selection	D02
C00105	Buy more advertising than misinformation creators	M009 - Dilution	Shift influence and algorithms by posting more adverts into spaces than misinformation creators.	Commercial advertising can be regarded as a form of legitimate expression. Caution is needed in some jurisdictions if there are campaign limits which may have an effect	TA07 Channel Selection	D03

				on the number of political ads a campaign may purchase.		
C00109	Dampen Emotional Reaction	M001 - Resilience	Reduce emotional responses to misinformation through calming messages, etc.	Not clear what this refers to or what actor would take this action. Could be calming replies to inflammatory or emotionally charged posts. Provided the action is taken transparently such counter-speech is an exercise in free speech.	TA09 Exposure	D03
C00111	Reduce polarisation by connecting and presenting sympathetic renditions of opposite views	M001 - Resilience		This appears to be a form of mediation or conflict resolution or conflict avoidance. Provided it is done transparently and the intent behind it is clear such counter-speech is an exercise in free speech.	TA01 Strategic Planning	D04
C00114	Don't engage with payloads	M004 - Friction	Stop passing on misinformation	In most democracies, every user of social media has the right to decide whether to pass on information, regardless of its veracity. Taking the time to ascertain the veracity and authenticity of information before engaging with it is a best practice to be encouraged.	TA08 Pump Priming	D02
C00115	Expose actor and intentions	M003 - Daylight	Debunk misinformation creators and posters.	If conducted by other users of the platform then this is simply an exercise of free speech. Similarly, if conducted by the media. However, if carried out by the platforms themselves, then the criteria for classifying content as misinformation needs to be transparent, clear, and consistently applied, since any judgment of this kind is subject to human or algorithmic bias.	TA08 Pump Priming	D02
C00125	Prebunking	M001 - Resilience	Produce material in advance of misinformation incidents, by anticipating the narratives used in them, and debunking them.	When carried out transparently and with a clearly stated intent, this is simply an exercise in free speech.	TA09 Exposure	D03

C00120	Open dialogue about design of platforms to produce different outcomes	M007 - Metatechnique	Redesign platforms and algorithms to reduce the effectiveness of disinformation	Encouraging platform designers to create platforms which are inhospitable to disinformation is a laudable goal in any democracy which depends upon a shared body of facts to make effective decisions. This includes the design of socio-technical systems to ensure that incentive structures do not encourage disinformation, conspiracy theories and online harms.	TA08 Pump Priming	D07
C00121	Tool transparency and literacy for channels people follow.	M001 - Resilience	Make algorithms in platforms explainable, and visible to people using those platforms.	Transparency and trust go together. Transparency and accountability also go together. Without transparency, algorithms can be harmful, or at a minimum, biased.	TA08 Pump Priming	D07
C00124	Don't feed the trolls	M004 - Friction	Don't engage with individuals relaying misinformation.	This is good general advice to social media users	TA09 Exposure	D03
C00211	Use humorous counter-narratives	M010 - Countermessaging		Humor can be very effective. Used in counter-narratives or counter-speech this is an application of freedom of expression.	TA09 Exposure	D03
C00130	Mentorship: elders, youth, credit. Learn vicariously.	M001 - Resilience	Train local influencers in countering misinformation.	Provided training and mentoring is carried out by civil society and not government and that such training and mentoring includes an education on the importance of freedom of speech, this is an effective way to build societal resilience to disinformation and online harm.	TA05 Microtargeting	D07
C00136	Microtarget most likely targets then send them	M010 - Countermessaging	Find communities likely to be targetted by misinformation campaigns, and send them countermessages or pointers to information sources.	When carried out transparently any type of counter-messaging or counter-speech is an exercise in freedom of expression.	TA08 Pump Priming	D03

	countermessages					
C00156	Better tell your country or organisation story	M010 - Countermessaging	Civil engagement activities conducted on the part of EFP forces. NATO should likewise provide support and training, where needed, to local public affairs and other communication personnel. Local government and military public affairs personnel can play their part in creating and disseminating entertaining and sharable content that supports the EFP mission.	Such public affairs or public diplomacy initiatives when conducted truthfully constitute an ethical means of interacting with the public sphere. The specific example refers to NATO's Enhanced Forward Presence which defends Eastern European states from possible incursion by Russia. Public diplomacy initiatives are vital to maintaining local support and to thwarting Kremlin active measures involving dezinformatsiya.	TA02 Objective Planning	D03
C00159	Have a disinformation response plan	M007 - Metatechnique	e.g. Create a campaign plan and toolkit for competition short of armed conflict (this used to be called “the grey zone”). The campaign plan should account for own vulnerabilities and strengths, and not over-rely on any one tool of statecraft or line of effort. It will identify and employ a broad spectrum of national power to deter, compete, and counter (where necessary) other countries’ approaches, and will include understanding of own capabilities, capabilities of disinformation creators, and international standards of conduct to compete in, shrink the size, and ultimately deter use of competition short of armed conflict.	While the example given is for a national plan, increasingly corporations and non-profits are advised to put disinformation response plans into place, so that they know what to do and who will do it, if and when they are subjected to a malign influence or smear campaign affecting their organization or brand.	TA01 Strategic Planning	D03

C00170	elevate information as a critical domain of statecraft	M007 - Metatechnique	Shift from reactive to proactive response, with priority on sharing relevant information with the public and mobilising private-sector engagement. Recent advances in data-driven technologies have elevated information as a source of power to influence the political and economic environment, to foster economic growth, to enable a decision-making advantage over competitors, and to communicate securely and quickly.	Information has long been recognized as an important element of national power. See, for example, Power and Interdependence in the Information Age on JSTOR. In the US, it is incorporated into models of statecraft such as DIME (Diplomatic, Information, Military, Economic) and PMESII (Political, Military, Economic, Social, Information, Infrastructure).	TA01 Strategic Planning	D03
C00174	Create a healthier news environment	M007 - Metatechnique, M002 - Diversion	Free and fair press: create bipartisan, patriotic commitment to press freedom. Note difference between news and editorialising. Build alternative news sources: create alternative local-language news sources to counter local-language propaganda outlets. Delegitimize the 24 hour news cycle. includes Provide an alternative to disinformation content by expanding and improving local content: Develop content that can displace geopolitically-motivated narratives in the entire media environment, both new and old media alike.	When news outlets are transparent and follow journalistic standards, they contribute to a healthy and vibrant democracy. The challenge is dealing with the headwinds of the rapidly changing political economy of the news created by advances in technology, especially the struggles of small, local news outlets to survive in a market that tends to evolve towards mega corporations. Governments can play a role here by providing financial support for independent media. Indeed, this is one of the elements of the "Resilience Building" pillar of the European Commission's FIMI Toolbox. See 2nd EEAS Report on Foreign Information Manipulation and Interference Threats | EEAS (europa.eu).	TA01 Strategic Planning	D02
C00182	Redirection / malware detection/ remediation	M005 - Removal	Detect redirection or malware, then quarantine or delete.	Cyber criminals use URL redirection to direct internet users to malware or phishing sites. Website scanners and web application firewalls can detect malicious URL redirects.	TA09 Exposure	D02

C00184	Media exposure	M003 - Daylight	highlight misinformation activities and actors in media	Some news services (e.g. BBC Verify) and many NGOs (e.g. Atlantic Council DFRLab) specialize in exposing misinformation actors and activities. This is an exercise in freedom of speech.	TA08 Pump Priming	D04
C00188	Newsroom/Journalist training to counter influence moves	M001 - Resilience	Includes SEO influence. Includes promotion of a “higher standard of journalism”: journalism training “would be helpful, especially for the online community. Includes Strengthen local media: Improve effectiveness of local media outlets.	Training newsrooms and journalists on how to detect online influence activities and promoting best practices on how to verify and report (or not report) on these activities is a vital component of societal resilience, given that news media and journalists are often direct targets of or "useful idiots" for these campaigns.	TA08 Pump Priming	D03
C00190	open engagement with civil society	M001 - Resilience	Government open engagement with civil society as an independent check on government action and messaging. Government seeks to coordinate and synchronise narrative themes with allies and partners while calibrating action in cases where elements in these countries may have been co-opted by competitor nations. Includes “fight in the light”: Use leadership in the arts, entertainment, and media to highlight and build on fundamental tenets of democracy.	Transparency is the key principle that differentiates democracies from totalitarian regimes and open discussions between democratic governments and civil society actors is to be encouraged, including discussions that promote democratic values and decry totalitarian values ("fight in the light" - see Training to Fight in the Light | Brennan Center for Justice). The caveat is that government must not abuse its power to unduly influence the conversation or "abridge" the freedom of speech of the people.	TA01 Strategic Planning	D03
C00200	Respected figure (influencer) disavows misinfo	M010 - Countermessaging	FIXIT: standardise language used for influencer/ respected figure.	The disavowal of misinformation is valid counter-speech and consistent with First Amendment values of free speech.	TA09 Exposure	D03
C00212	build public resilience by making civil society more vibrant	M001 - Resilience	Increase public service experience, and support wider civics and history education.	Expanding access to education in civics and history and encouraging participation in public service are great ways of building resilience to disinformation and foreign manipulation	TA01 Strategic Planning	D03

C00219	Add metadata to content that's out of the control of disinformation creators	M003 - Daylight	Steganography. Adding date, signatures etc to stop issue of photo relabelling etc.	Adding metadata to content when it is created or altered is a way to inform users about its provenance. Specifically, watermarking, steganography, or cryptography can be used to establish content authenticity and prevent copyright abuse or the tampering of content for use in disinformation and influence campaigns. Adobe's Content Authenticity Initiative is a cross-industry, open-source initiative to promote a standard approach to adding tamper-evident provenance to all types of digital content, starting with photos, video, and documents (Content Authenticity Initiative). The approach is compatible with the Coalition for Content Provenance and Authenticity, or C2PA, which has gained traction recently as one way of distinguishing AI-generated from human-generated content (The inside scoop on watermarking and content authentication | MIT Technology Review).	TA06 Develop Content	D04
C00220	Develop a monitoring and intelligence plan	M007 - Metatechnique	Create a plan for misinformation and disinformation response, before it's needed. Include connections / contacts needed, expected counteremessages etc.	Governments and corporations have begun to develop disinformation response plans which are similar in nature to incident response plans that deal with cyber-attacks. Monitoring and intelligence collection plans are also part of the planning efforts behind disinformation risk management.	TA01 Strategic Planning	D03
C00221	Run a disinformation red team, and design mitigation factors	M007 - Metatechnique	Include PACE plans - Primary, Alternate, Contingency, Emergency	Red teaming is a great way of learning about potential threats and of developing ways to mitigate the threats, including ensuring that an effective communications plan is in place. Primary, alternate, contingency and emergency (PACE) is a methodology used to build a communication plan.	TA01 Strategic Planning	D03

C00222	Tabletop simulations	M007 - Metatechnique	Simulate misinformation and disinformation campaigns, and responses to them, before campaigns happen.	Tabletop exercises are an excellent way to teach decision-makers about potential threats and how to deal with them.	TA02 Objective Planning	D03
C00223	Strengthen Trust in social media platforms	M001 - Resilience	Improve trust in the misinformation responses from social media and other platforms. Examples include creating greater transparancy on their actions and algorithms.	Transparency is one of the key ingredients of trust. As social media platforms are the guardians of today's public sphere, even though they are private entities that are not beholden to the same laws and regulations that governments are, and so there is an expectation amongst the public that they are transparent about how they identify and respond to disinformation and misinformation. As much of the detection and response is automated, the transparency concerning algorithms is very important. Particularly when social media companies take enforcement measures which appear to limit the free speech rights of some users to mitigate the perceived harm to other users, social media companies need to be forthcoming about their policies and their actions and provide recourse for appeals.	TA01 Strategic Planning	D03

Table 4 Counters which are potentially problematic

disarm_id	name	metatechnique	summary	ethical and legal considerations	tactic	response type
C00012	Platform regulation	M007 - Metatechnique	Empower existing regulators to govern social media. Also covers Destroy. Includes: Include the role of social media in the regulatory framework for media. The U.S. approach will need to be carefully crafted to protect First Amendment principles, create needed transparency, ensure liability, and impose costs for noncompliance. Includes Create policy that makes social media police disinformation. Includes: Use fraud legislation to clean up social media	This is controversial in the US where Section 230 is the current federal law stating that websites are not liable for third party content and where efforts to abolish Section 230 are largely partisan. In Europe VLOPs are regulated through the Digital Services Act and Digital Markets Act, but it is likely that these statutes will be challenged in court by the large tech companies. In the US there are a handful of cases including child pornography and imminent physical danger caused by incitement to violence in which freedom from harm overrides freedom of speech. But rapid advancements in social media technology have enabled a rise in harms such as cyber bullying, misogyny, cyber stalking, promotion of self-harm, doxing, cancel culture, and a concomitant rise in teenage suicides, mental health problems, as well as a chilling effect among targeted communities which end up exercising self-censorship rather than face the ire of the online mob. Laws protecting online users from digital harm may take a long time to catch up. Smaller countries may act more quickly e.g. the United Kingdom has passed the Online Safety Bill, motivated in part by an intense debate about how to balance free expression versus harmful content targeted at children. The law requires digital platforms to screen proactively for illegal content such as child pornography or promotion of self-harm.	TA01 Strategic Planning	D02

C00013	Rating framework for news	M006 - Scoring	This is "strategic innoculation", raising the standards of what people expect in terms of evidence when consuming news. Example: journalistic ethics, or journalistic licencing body. Include full transcripts, link source, add items.	This includes services such as Newsguard which can be very useful when available as an optional add-on to an internet browser or a social media platform or as a feature that can be enabled in a user's profile. Enforcing the use of such rating frameworks by consumers, however, would be problematic. Internet users have a right to view their favorite news sources unencumbered by ratings services and fact-checkers, even when those sources are spreading untruths. If a platform performs fact-checking of news and then downranks and/or labels content deemed false, then ideally it provides its users with the choice to opt out e.g. Facebook user profiles include a setting to opt-out of the downranking of fact-checked content. See Following Musk's lead, Youtube and Facebook are giving up on policing conspiracies - The Washington Post.	TA01 Strategic Planning	D02
C00017	Repair broken social connections	M010 - Countermessaging	For example, use a media campaign to promote in-group to out-group in person communication / activities . Technique could be in terms of forcing a reality-check by talking to people instead of reading about bogeymen.	This feels like social engineering. If done transparently it could be ok.	TA01 Strategic Planning	D03
C00019	Reduce effect of division-enablers	M003 - Daylight	includes Promote constructive communication by shaming division-enablers, and Promote playbooks to call out division-enablers	Naming and shaming at a national diplomatic level e.g. through the UN may be acceptable but when done domestically this can border on defamation. Provided this is done respectfully without ad hominem attacks, calling out others for divisive behaviors online and proposing more constructive approaches should be acceptable.	TA01 Strategic Planning	D03

C00024	Promote healthy narratives	M001 - Resilience	Includes promoting constructive narratives i.e. not polarising (e.g. pro-life, pro-choice, pro-USA). Includes promoting identity neutral narratives.	Problematic if done clandestinely. As long as this is done transparently it is ok. It might be effective if the party doing the promotion has a position of trust within the community	TA01 Strategic Planning	D04
C00031	Dilute the core narrative - create multiple permutations, target / amplify	M009 - Dilution	Create competing narratives. Included "Facilitate State Propaganda" as diluting the narrative could have an effect on the pro-state narrative used by volunteers, or lower their involvement.	Creating multiple counter-narratives is potentially problematic if they either contain dis or misinformation or involve an attempt to flood the information environment. Both of these are manipulative in nature. If, on the other hand, this action involves countering a simplistic narrative by exploring different nuances of the truth rather than spreading falsehoods or bombarding the information environment, then this can be democratically acceptable.	TA02 Objective Planning	D03
C00032	Hijack content and link to truth- based info	M002 - Diversion	Link to platform	It depends upon how this is carried out. "Hijacking" sounds manipulative e.g. if a hashtag that clearly has one meaning is "hijacked" to promote an alternative meaning without transparency around the intent behind the "hijacking" then this is manipulative, albeit not illegal or violative of most terms of servce.	TA06 Develop Content	D03
C00034	Create more friction at account creation	M004 - Friction	Counters fake account	This depends on what the friction entails. If Know Your Customer regulations are enforced and anonymity disallowed, then users promoting human rights or democratic principles within an oppressive political system will be placed in danger. There may not be a one size fits all answer.	TA15 - Establish Social Assets	D04
C00067	Denigrate the recipient/ project (of online funding)	M013 - Targeting	Reduce the credibility of groups behind misinformation-linked funding campaigns.	Exposing the funding behind disinformation efforts, especially when such funding is covert, can be a valuable service, as long as it is done properly. However, attempting to denigrate the recipient, project, or funder,	TA15 Establish Social Assets	D03

				only leads to a race to the bottom and may be counter-productive or create the opposite of the desired effect		
C00044	Keep people from posting to social media immediately	M004 - Friction	Platforms can introduce friction to slow down activities, force a small delay between posts, or replies to posts.	Who decides whose posts or what types of posts are slowed down? Any type of throttling back is potentially problematic if it is done selectively, unless the criteria are transparent for all to see.	TA15 - Establish Social Assets	D03
C00046	Marginalise and discredit extremist groups	M013 - Targeting	Reduce the credibility of extremist groups posting misinformation.	It is not clear exactly what this counter entails. The biggest challenge is deciding which groups are "extremist" and ensuring the consistent and transparent application of a uniform classification standard. If such a classification were based clearly on law enforcement designations, then this may be acceptable, otherwise this could be highly controversial.	TA15 - Establish Social Assets	D04
C00048	Name and Shame Influencers	M003 - Daylight	Think about the different levels: individual vs state-sponsored account. Includes “call them out” and “name and shame”. Identify social media accounts as sources of propaganda—“calling them out”—might be helpful to prevent the spread of their message to audiences that otherwise would consider them factual. Identify, monitor, and, if necessary, target externally-based nonattributed social media accounts. Impact of and Dealing with Trolls - "Chatham House has observed that	Calling out sources of propaganda or misinformation is an exercise in counter-speech and is consistent with freedom of expression. When this is done at an international level in diplomatic circles it is often highly appropriate. Calling out trolls and inauthentic accounts can also be a service to a domestic online community. However, great care has to be taken that this is not done in a toxic manner or involves ad hominem attacks that quickly spiral into online street brawls or, even worse, catalyse a mob mentality. There is a line at which	TA15 - Establish Social Assets	D07

			trolls also sometimes function as decoys, as a way of “keeping the infantry busy” that “aims to wear down the other side” (Lough et al., 2014). Another type of troll involves “false accounts posing as authoritative information sources on social media”.	counter-speech can devolve into cancel culture and mob harassment.		
C00056	Encourage people to leave social media	M004 - Friction	Encourage people to leave spcial media. We don't expect this to work	Any type of campaign that might encourage a boycott or similar of a platform or service may encounter legal problems. On the other hand, providing reviews of different platforms in terms of features, pricing etc. to assist consumers in making the most appropriate choice for them would be unproblematic.	TA15 Establish Social Assets	D02
C00058	Report crowdfunder as violator	M005 - Removal	counters crowdfunding. Includes ‘Expose online funding as fake”.	Not clear why crowdfunding is being portrayed as a bad thing or what it is violating	TA15 - Establish Social Assets	D02
C00065	Reduce political targeting	M005 - Removal	Includes “ban political micro targeting” and “ban political ads”	Any type of ban or removal is potentially violative of freedom of speech. Clearly, commercial platforms can decide on their own terms and conditions within the laws in which they operate. Twitter initially banned political ads, while Facebook argued against fact-checking political ads and in favor of politicians' speech being newsworthy. Elon Musk then reversed Twitter's policy and reversed the ban on political ads.	TA05 Microtargeting	D03
C00066	Co-opt a hashtag and drown it out (hijack it back)	M009 - Dilution	Flood a disinformation-related hashtag with other content.	Any type of flooding of the information environment is manipulative of public discourse. Nevertheless, there may be specific circumstances in which malign influence operations are using automation to spread harmful information (e.g. false information on election polling station	TA05 Microtargeting	D03

				locations or times). In such circumstances, an automated response that promotes accurate information may be warranted if it is done transparently.		
C00080	Create competing narrative	M002 - Diversion	Create counternarratives, or narratives that compete in the same spaces as misinformation narratives. Could also be degrade	Creating and propagating counternarratives is an exercise in freedom of expression but can be problematic if the counter-speech is carried out covertly. Any actions that involve "degrading" are potentially problematic from a democratic point of view.	TA06 Develop Content	D03
C00071	Block source of pollution	M005 - Removal	Block websites, accounts, groups etc connected to misinformation and other information pollution.	The acceptability of this action highly depends on circumstances. Corporations routinely use web filtering technologies to filter out categories of sites they do not want their employees to view. Parents filter what their children can see. It gets problematic, however, in terms of freedom of thought and freedom of speech, when technology platforms or internet service providers use such technologies without transparency or without providing their users the ability to customize what is filtered.	TA06 Develop Content	D02
C00074	Identify and delete or rate limit identical content	M012 - Cleaning	C00000	Platforms are allowed to implement their own algorithms which rate limit content and introduce their own terms and conditions regarding the posting of duplicate content. Removal of content is inherently problematic when it comes to freedom of speech, so many platforms will instead choose to rate-limit potentially problematic content. When the factors involved in such rate-limiting algorithms are not explained or the algorithms are opaque, then there is a risk that the rights of some groups or individuals are being infringed. The key is transparency. Legally, however, it may depend upon jurisdiction as to what the	TA06 Develop Content	D02

				platform is allowed to do: in the US, commercial platforms enjoy editorial discretion derived from their own First Amendment Right to freedom of the press.		
C00077	Active defence: run TA15 "develop people” - not recommended	M013 - Targeting	Develop networks of communities and influencers around counter-misinformation. Match them to misinformation creators	If done transparently and such influencers made it very clear what their intentions, motivation, and funding sources are, then I believe this to be an exercise in free speech and in participating in the marketplace of ideas. If done clandestinely, however, this feels like social engineering.	TA15 - Establish Social Assets	D03
C00078	Change Search Algorithms for Disinformation Content	M002 - Diversion	Includes “change image search algorithms for hate groups and extremists” and “Change search algorithms for hate and extremist queries to show content sympathetic to opposite side”	This approach has been employed to counter violent extremism by Moonshot CVE in partnership with Google Jigsaw and the Anti-Defamation League in a program named the Redirect Method. The Search for Extremism: Deploying the Redirect Method | The Washington Institute. Search engine platforms have a lot of latitude about how they implement their algorithms. In this case there was no censorship, just alternative viewpoints. Still, tech platforms need to be transparent about such efforts and provide researchers with data on them or there is a risk that certain groups will be unintentionally and unknowingly impacted.	TA06 Develop Content	D03
C00082	Ground truthing as automated response to pollution	M010 - Countermessaging	Also inoculation.	This seems to involve automated responses that include empirical evidence so presumably there would need to be automated detection and classification of what constitutes "pollution" in the first place and then a process for matching that with empirical evidence. The danger of exacerbating the situation with a flood of false positives is very real here.	TA06 Develop Content	D03

C00085	Mute content	M003 - Daylight	Rate-limit disinformation content. Reduces its effects, whilst not running afoul of censorship concerns. Online archives of content (archives of websites, social media profiles, media, copies of published advertisements; or archives of comments attributed to bad actors, as well as anonymized metadata about users who interacted with them and analysis of the effect) is useful for intelligence analysis and public transparency, but will need similar muting or tagging/ shaming as associated with bad actors.	This counter appears to conflate the tagging or flagging of content as disinformation (daylight) with the rate-limiting of such content (friction). Rate-limiting is certainly preferable to removal or censorship and is the preferred choice of some tech platforms for dealing with content that is not illegal, but which is deemed potentially problematic. Nevertheless, unless platform policies make it very clear what their policies are regarding classifying and handling problematic content this is subject to abuse by and the bias of the authors of the policy and associated algorithms.	TA06 Develop Content	D03
C00092	Establish a truth teller reputation score for influencers	M006 - Scoring	Includes "Establish a truth teller reputation score for influencers” and “Reputation scores for social media users”. Influencers are individuals or accounts with many followers.	Reputation scores for influencers could be useful if available as a separate service or even as an optional add-on to an internet browser or a social media platform or as a feature that can be enabled in a user's profile. Enforcing the use of a rating framework by consumers, however, would be problematic. Internet users have a right to consume content from their favorite influencers unencumbered by ratings services and fact-checkers, even when those sources are spreading untruths. If a platform performs fact-checking of news and then downranks and/or labels content deemed false, then ideally it provides its users with the choice to opt out e.g. Facebook user profiles include a setting to opt-out of the downranking of fact-checked content. See Following Musk’s lead, Youtube and Facebook are giving up on policing conspiracies - The Washington Post. In reality digital platforms in most democratic	TA02 Objective Planning	D07

				jurisdictions have considerable latitude about which features they provide and how visible these features are.		
C00093	Influencer code of conduct	M001 - Resilience	Establish tailored code of conduct for individuals with many followers. Can be platform code of conduct; can also be community code.	Any such code of conduct would have to be aspirational only and should not be mandated, so that the influencer's right to freedom of expression are not abridged	TA15 - Establish Social Assets	D07
C00096	Strengthen institutions that are always truth tellers	M006 - Scoring	Increase credibility, visibility, and reach of positive influencers in the information space.	The validity of this action depends largely on the actor taking it. Commercial platforms typically have a lot of leeway to prioritize certain sources in their ranking algorithms or even to boost the visibility and reach of these sources. In the US, for example, this freedom comes from the platforms' own editorial discretion (freedom of the press) and their own free speech rights. Governments intervening in the information environment to favor one type of source over another may be highly problematic. In the US such action can quickly fall foul of First Amendment jurisprudence, except when it comes to assertions of fact (such as where and when to vote) that are clearly within the Government's stated responsibilities.	TA01 Strategic Planning	D07
C00098	Revocation of allowlisted or "verified" status	M004 - Friction	remove blue checkmarks etc from known misinformation accounts.	Revocation of verified status based on a platform making decisions about what is or isn't misinformation is problematic. Where content is clearly illegal or a user's behavior clearly violates the platform's terms and conditions, then revoking verified status may be an appropriate response, but in some geographies, such as the United States, citizens enjoy a constitutional right to lie, so revocation purely on the basis of	TA07 Channel Selection	D02

				untruthful content would feel like a violation of their First Amendment rights, even if the platform, as a private actor, has considerable latitude to decide what types of content can and cannot be published on their platform, given the platform's own rights to freedom of speech and freedom of the press.		
C00100	Hashtag jacking	M002 - Diversion	Post large volumes of unrelated content on known misinformation hashtags	It depends upon how this is carried out: "jacking" sounds manipulative e.g. if a hashtag that clearly has one meaning is hijacked to promote an alternative meaning without transparency around the intent behind the "hijacking" then this is manipulative, albeit not necessarily illegal or violative of terms of service. The Chinese Communist Party uses this technique to drown out human rights demonstrators.	TA08 Pump Priming	D03
C00101	Create friction by rate-limiting engagement	M004 - Friction	Create participant friction. Includes Make repeat voting hard, and throttle number of forwards.	Social media platforms typically set rate limits for engagement (how frequently you can like, share, repost etc.) to weed out spam and inauthentic activity. For power users this may cause some frustration and may impede the flow of ideas. A careful balance should be struck. When terms and conditions or community guidelines are violated, engagement may be restricted or blocked: that is where concerns regarding freedom of expression are highest. Nevertheless, depending on jurisdiction, commercial platforms may enjoy considerable latitude to rate-limit or even block content: this is the case in the US where the platforms enjoy freedom of speech and freedom of the press.	TA07 Channel Selection	D04

C00107	Content moderation	M006 - Scoring, M005 - Removal	includes social media content take-downs, e.g. facebook or Twitter content take-downs	Every platform publishes community guidelines or terms and conditions that include what types of content are not allowed. These policies need to be transparent and enforcement of them consistently applied. Moreover, there needs to be transparency about content moderation actions that platforms have taken, usually in so-called transparency reports, for the purposes of accountability. And ideally these policies and actions are reviewed by an independent oversight body. When these elements are not in place or they are abused, then content moderation can turn into censorship. At the end of the day, however, commercial platforms may enjoy the freedom to decide how to handle content on their platform. This is the case in the US, for example, where platforms have free speech and free press rights of their own.	TA06 Develop Content	D02
C00118	Repurpose images with new text	M010 - Countermessaging	Add countermessage text to iamges used in misinformation incidents.	In social media, replies or reposts with comments by users constitute counter-speech, which is an exercise of free speech. However, if the platform itself labels images as misinformation e.g. based on fact-checking or AI detection, the rationale and the policy should be clear and consistently applied. Some users may want to know, others may not. For example, if the content originates from an agent of a foreign power, users in certain countries may feel they have a right to know this, while other users may not care. For those users who do not care, unwanted labels can be construed as an impediment to the free flow of ideas, even if those ideas contain untruths. It may come down to societal norms as to whether	TA08 Pump Priming	D04

				labeling of platform-designated misinformation is acceptable, and if so, for what types of misinformation. If a platform performs fact-checking of images and then labels the image, then ideally it provides its users with the choice to opt out of such labeling. At the end of the day, however, commercial platforms may enjoy the freedom to decide whether and what to label. This is the case in the US, for example, where platforms have free speech and free press rights of their own.		
C00113	Debunk and defuse a fake expert / credentials.	M003 - Daylight	Debunk fake experts, their credentials, and potentially also their audience quality	Debunking fake credentials is an exercise of free speech but impugning the character of an audience is a type of ad hominem attack and does not comport with civil discourse.	TA08 Pump Priming	D02
C00116	Provide proof of involvement	M003 - Daylight	Build and post information about groups etc's involvement in misinformation incidents.	When covert disinformation causes harm, one approach is to publicly call the actor behind it to account. However, this is fraught with the risk of misattribution and innocent parties being accused of something they had nothing to do with, or worse still, it can devolve into a witch hunt and mob harassment or vilification. A better approach in the event that the actor or group behind misinformation has violated a law or a policy is to report the actor or group to law enforcement or platform administrators as appropriate.	TA08 Pump Priming	D02
C00117	Downgrade / de-amplify so message is seen by fewer people	M010 - Countermessaging	Label promote counter to disinformation	Platforms may decide to downgrade messages or posts which are suspicious or that appear to violate their policies instead of blocking or removing them. This is better than blocking or removing but still relies on an accurate classification of the message or post as disinformation in the first place. The classification is subject to human or algorithmic bias. Furthermore, if the	TA08 Pump Priming	D04

				classification is made purely on the basis of the veracity of the content, this can be problematic in countries with strict norms of freedom of expression such as the US, where citizens largely enjoy the right to lie. Nevertheless, social media platforms are not government owned. They are owned by private actors who enjoy editorial discretion in the US because of First Amendment freedom of the press, so they have a lot of latitude when it comes to handling content on their platform.		
C00119	Engage payload and debunk.	M010 - Countermessaging	debunk misinformation content. Provide link to facts.	When regular social media users respond e.g. using a comment or reply then this is valid counter-speech consistent with the marketplace of ideas. When social media platforms do the same thing by labeling content this can be viewed as an obstruction in the flow of ideas. If a platform performs fact-checking and then downranks and/or labels content deemed false, then ideally it provides its users with the choice to opt out e.g. Facebook user profiles include a setting to opt-out of the downranking of fact-checked content. See Following Musk's lead, Youtube and Facebook are giving up on policing conspiracies - The Washington Post. Each platform has its own policies on fact-checking and labeling of content, and, in most jurisdictions, is free to do so in accordance with the platform's own editorial discretion and free speech rights. Many platforms work with independent fact-checkers. Some platforms do very little or no fact-checking. This is largely up to the platforms in the U.S. but in Europe there are stricter laws regarding potential online	TA08 Pump Priming	D07

				harms and platforms are responsible for screening for content that may be illegal.		
C00122	Content moderation	M004 - Friction	Beware: content moderation misused becomes censorship.	Every platform publishes community guidelines or terms and conditions that include what types of content are not allowed. These policies should be transparent and enforcement of them consistently applied. Moreover, there should be transparency about content moderation actions that platforms have taken, usually in so-called transparency reports, for the purposes of accountability. And ideally these policies and actions are reviewed by an independent oversight body. When these elements are not in place or they are abused, then content moderation can turn into censorship. At the end of the day, however, commercial platforms may enjoy the freedom to decide which content they want to moderate. This is the case in the US, for example, where platforms have free speech and free press rights of their own.	TA09 Exposure	D02
C00123	Remove or rate limit botnets	M004 - Friction	reduce the visibility of known botnets online.	While bots, or automated software programs, may or may not be malicious, botnets, defined as a group of computers infected with malicious software and remotely controlled without their owners' knowledge, are always malicious. In most jurisdictions operating a botnet is illegal, but sinkholes and takedowns need to take account of laws against illegal access to or intrusion of other people's computers. For example, in the United States this includes the Computer Fraud and Abuse Act and the Fourth Amendment. Therefore, any such interventions should be carried out by law	TA09 Exposure	D03

				enforcement or authorized personnel only. Throttling involves blocking traffic from specific users who exceed certain thresholds: in certain jurisdictions, such as the EU, this may be illegal due to laws upholding open internet access, while in others, such as the US, this may be legal due to the repeal of net neutrality, provided the service provider's contract spells out the circumstances in which throttling is undertaken. Rate-limiting, on the other hand, is a defensive measure that clients are generally always permitted to take to protect themselves.		
C00126	Social media amber alert	M003 - Daylight	Create an alert system around disinformation and misinformation artefacts, narratives, and incidents	If a social media platform issues alerts pertaining to disinformation or misinformation e.g. based on fact-checking or AI detection, the rationale and the policy should be clear and consistently applied. Some users may want to know, others may not. For those users who do not care, obtrusive alerts can be construed as an impediment to the free flow of ideas, even if those ideas contain untruths. Implementing alerts unobtrusively, where the alerts are available to those who want to view them, may be a good compromise. Alternatively, providing users with the choice of whether they want to see alerts through profile settings strikes a good balance. It may come down to societal norms as to whether obtrusive disinformation or misinformation alerts are acceptable, and if so, for what types of misinformation. At the end of the day, however, commercial platforms may enjoy the freedom to decide whether and what to label. This is the case in the US, for	TA09 Exposure	D03

				example, where platforms have free speech and free press rights of their own.		
C00128	Create friction by marking content with ridicule or other "decelerants"	M009 - Dilution	Repost or comment on misinformation artefacts, using ridicule or other content to reduce the likelihood of reposting.	Organic reposting or commenting by regular users is an exercise in freedom of expression. When carried in automated fashion this may be problematic, especially if it is not done transparently. When carried out by the platform itself this would be highly problematic, as the platform would be deciding by itself what is and what is not disinformation or misinformation and then interfering in public discourse. At the end of the day, however, commercial platforms may enjoy the freedom to decide whether and what to mark or decelerate. This is the case in the US, for example, where platforms have free speech and free press rights of their own.	TA09 Exposure	D03
C00131	Seize and analyse botnet servers	M005 - Removal	Take botnet servers offline by seizing them.	While bots, or automated software programs, may or may not be malicious, botnets, defined as a group of computers infected with malicious software and remotely controlled without their owners' knowledge, are always malicious. In most jurisdictions operating a botnet is illegal, but sinkholes and takedowns need to take account of laws against illegal access to or intrusion of other people's computers. For example, in the United States this includes	TA11 Persistence	D02

				the Computer Fraud and Abuse Act and the Fourth Amendment. Therefore, any such interventions should be carried out by law enforcement or authorized personnel only.		
C00133	Deplatform Account*	M005 - Removal	Note: Similar to Deplatform People but less generic. Perhaps both should be left.	(1) This is an extreme measure that social media platforms can take when users violate their Terms of Service. The argument against deplatforming is that it can "draw attention to suppressed materials (Streisand effect), harden the conviction of the followers, and put social media companies in the position of an arbiter of speech". The argument for deplatforming is that it "detoxes both subspaces (such as subreddits) as well as platforms more generally, produces a decline in audience and drives extreme voices to spaces that have less oxygen-giving capacity, thereby containing their impact". See Richard Rogers's article at Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media - Richard Rogers, 2020 (sagepub.com). Deplatformed users and their followers typically migrate to other platforms with less restrictive Terms of Service. (2) **In the United States,** social media companies enjoy considerable freedom when it comes to deciding to suspend or remove users' accounts. This freedom comes from their free speech and free press rights enshrined in the First Amendment: "Congress shall make no law respecting an establishment of religion or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably	TA15 - Establish Social Assets	D03

				to assemble, and to petition the Government for a redress of grievances". While the government is prohibited from abridging the freedom of speech of US citizens, social media platforms are themselves private entities that enjoy freedom of speech and freedom of the press: as such they enjoy broad editorial discretion about what gets published on their platform. This right is reinforced by Section 230, which states that they are not liable for third party content on their platform, including choosing not to publish such content. In fact, it is usually stated clearly in a platform's Terms of Service that the platform can decide to remove content or a user account at its own discretion (i.e. it can do whatever it wants). Nevertheless, such power in the hands of large quasi-monopolistic corporations can be abused. At the time of writing (2023), statutes have been passed by Florida and Texas which attempt to regulate social media platforms' ability to deplatform accounts. Some academics regard these laws as largely performative and political in nature, believing that they will ultimately be struck down by the US Supreme Court. See interview with Professor Eric Goldman at Online CLE & MCLE | The Law of Deplatforming (talksonlaw.com). (3) **In Europe**, the legal framework is quite different than in the United States. Recital 22 of the Digital Services Act states that the "removal or disabling of access should be undertaken in the observance of the principle of freedom of expression" (Texts adopted - Digital Services Act ***I -		

				Thursday, 20 January 2022 (europa.eu)). The DSA protections of freedom of expression come in the form of procedural safeguards: when platforms disable an account based on a violation of their rules, they are required to provide a "clear and specific statement of reasons" for the decision, including the "facts and circumstances relied on in taking the decision", and "explanations as to why the information is considered to be incompatible" with their policy. Platforms are required to have due regard to the fundamental rights of users under the EU Charter of Fundamental Rights, including freedom of expression. See Deplatforming Politicians and the Implications for Europe – Global Digital Cultures.		
C00135	Deplatform message groups and/or message boards	M005 - Removal	Merged two rows here.	The ethical and democratic reasoning behind deplatforming message groups and/or message boards is similar to that of deplatforming accounts. See "Deplatform Account" above.	TA15 Establish Social Assets	D03
C00142	Platform adds warning label and decision point when sharing content	M004 - Friction	Includes “this has been disproved: do you want to forward it”. Includes ““Hey this story is old" popup when messaging with old URL” - this assumes that this technique is based on visits to an URL shortener or a captured news site that can publish a message of our choice. Includes “mark clickbait visually”.	If a platform adds warning labels pertaining to disinformation or misinformation e.g. based on fact-checking or AI detection, the rationale and the policy should be clear and consistently applied. Some users may want to know, others may not. For those users who do not care, obtrusive labels can be construed as an impediment to the free flow of ideas, even if those ideas contain untruths. Ideally the platform would offer a setting in a user's profile indicating whether or not the user wants to see fact-checking labels. But this may impose additional costs. Implementing labels unobtrusively, where the labels are available to those who want to view them, may be a good compromise. It	TA06 Develop Content	D04

				may come down to societal norms as to whether obtrusive disinformation or misinformation labels are acceptable, and if so, for what types of misinformation. At the end of the day, however, commercial platforms may enjoy the freedom to decide whether and what to label. This is the case in the US, for example, where platforms have free speech and free press rights of their own.		
C00143	(botnet) DMCA takedown requests to waste group time	M013 - Targeting	Use copyright infringement claims to remove videos etc.	Genuine requests for content removal based on infringement of the Digital Millenium Copyright Act are ethical. Frivolous requests based upon unsubstantiated claims of infringement are not.	TA11 Persistence	D04
C00147	Make amplification of social media posts expire (e.g. can't like/ retweet after n days)	M004 - Friction	Stop new community activity (likes, comments) on old social media posts.	Blocking features within social media platforms impedes the free flow of information in the marketplace of ideas. It may, however, be quite legal for social media platforms to take such actions, especially in the US where the platforms enjoy editorial discretion due in part to the free press clause of the First Amendment.	TA09 Exposure	D03
C00154	Ask media not to report false information	M005 - Removal	Train media to spot and respond to misinformation, and ask them not to post or transmit misinformation they've found.	Training or even asking journalists not to report on disinformation which has not broken out of its own echo chamber or misinformation which has so far had little impact or caused no harm is not really "removal": it would be more apt to call this "withholding". The purpose of this countermeasure is to ensure that the media does not give oxygen to dis- or misinformation that would otherwise die on the vine. Such advocacy needs to be completely free of any coercion. For those journalists who decide to report on dis- or misinformation, encouraging them to do so	TA08 Pump Priming	D02

				using so-called “truth sandwiches” may be appropriate. This involves covering the truth upfront in the story before covering the misinformation and then ending the story by presenting the truth again. This technique helps to avoid further spreading the dis- or misinformation. See How to serve up a tasty ‘truth sandwich?’ - Poynter. However, if the government is the one doing the asking, then the question of coercion may be unclear and this counter may be ethically problematic, unless there are national security implications or there are lives at stake. In democracies, freedom of the press is pivotal to holding the government accountable. In the US, this freedom was reaffirmed in 1971 in the Pentagon Papers case in which the judges quoted from two other decisions: “Any system of prior restraints of expression comes to this court bearing a heavy presumption against its constitutional validity” … the government “thus carries a heavy burden of showing justification for the imposition of such a restraint”...."The District Court for the Southern District of New York, in The New York Times case, and the District Court for the District of Columbia and the Court of Appeals for the District of Columbia Circuit, in The Washington Post case, held that the government had not met that burden. We agree." See New York Times Co. v. United States (The Pentagon Papers Case) | Constitution Center.		
C00160	find and train influencers	M001 - Resilience	Identify key influencers (e.g. use network analysis), then reach out to	When carried out transparently this is an effective way to promote accurate information and positive narratives	TA15 - Establish	D02

			identified users and offer support, through either training or resources.		Social Assets	
C00161	Coalition Building with stakeholders and Third-Party Inducements	M007 - Metatechnique	Advance coalitions across borders and sectors, spanning public and private, as well as foreign and domestic, divides. Improve mechanisms to collaborate, share information, and develop coordinated approaches with the private sector at home and allies and partners abroad.	Such coalitions as the OECD DIS/MIS Resource Hub can be highly effective but care must be taken that governments do not unduly influence civil society actors on what is acceptable or unacceptable speech, and any such coalitions should operate transparently to generate trust and ensure no abuse.	TA01 Strategic Planning	D07
C00165	Ensure integrity of official documents	M004 - Friction	e.g. for leaked legal documents, use court motions to limit future discovery actions	This action appears to refer to the use of protective orders during litigation. When used properly protective orders protect trade secrets or other privileged communication from public disclosure. See 22 CFR § 224.24 - Protective order. | Electronic Code of Federal Regulations (e-CFR) | US Law | LII / Legal Information Institute (cornell.edu). When used unethically protective orders can be abused by plaintiffs to extract a larger settlement from defendants who do not wish to see certain documents disclosed. See When the Bell Can't Be Unrung: Document Leaks and Protective Orders in Mass Tort Litigation (wne.edu).	TA06 Develop Content	D02
C00169	develop a creative content hub	M010 - Countermessaging	international donors will donate to a basket fund that will pay a committee of local experts who will, in turn, manage and distribute the money to Russian-language producers and broadcasters that pitch various projects.	A content hub is simply a curated ollection of content that offers a deep dive on a specific topic (What is a content hub? - Optimizely). Developing such a hub transparently is unproblematic. The example, given in the summary, however, appears to describe a covert operation to fund locally produced Russian language content favorable to those mounting the operation. Such activities should only be	TA02 Objective Planning	D03

				carried out by personnel authorized by the government.		
C00172	social media source removal	M005 - Removal	Removing accounts, pages, groups, e.g. facebook page removal	The considerations here are identical to those discussed under "Deplatform Account" above	TA15 Establish Social Assets	D02
C00176	Improve Coordination amongst stakeholders: public and private	M007 - Metatechnique	Coordinated disinformation challenges are increasingly multidisciplinary, there are few organisations within the national security structures that are equipped with the broad-spectrum capability to effectively counter large-scale conflict short of war tactics in real-time. Institutional hurdles currently impede diverse subject matter experts, hailing from outside of the traditional national security and foreign policy disciplines (e.g., physical science, engineering, media, legal, and economics fields), from contributing to the direct development of national security countermeasures to emerging conflict short of war threat vectors. A Cognitive Security Action Group (CSAG), akin to the Counterterrorism Security Group (CSG), could drive interagency alignment across equivalents of DHS, DoS, DoD, Intelligence Community, and other implementing agencies, in areas including strategic narrative, and the	The danger with public-private partnerships is that the government can use its power to coerce a commercial platform to promote or remove content on its platform. Called "jawboning", such action "is dangerous because it allows government officials to assume powers not granted to them by law" (Jawboning against Speech | Cato Institute). In a landmark case currently before the US Supreme Court, Murthy v. Missouri (formerly Missouri v. Biden), the plaintiffs argue that communication between Biden administration officials and social media platforms like Facebook and Twitter in recent years constituted a violation of the First Amendment. The defendants argue that the administration was taking responsible action to protect public health, safety and security when confronted with the challenges of a deadly pandemic and foreign attacks on elections (Appeals Court Rules White House Overstepped 1st Amendment on Social Media - The New York Times (nytimes.com)). The case will rest upon whether the plaintiffs can satisfy the "state action requirement" (state action requirement | Wex | US Law | LII / Legal Information Institute (cornell.edu)) by	TA01 Strategic Planning	D07

			nexus of cyber and information operations.	proving that the government's communication rose to the level of a concerted coercion campaign such that ensuing actions taken by the platform can be attributed to the state in violation of the First Amendment (Missouri v. Biden: An Opportunity to Clarify Messy First Amendment Doctrine | Knight First Amendment Institute (knightcolumbia.org)). The court has agreed to hear the case in the 2023-2024 term (Murthy v. Missouri - Wikipedia).		
C00178	Fill information voids with non-disinformation content	M009 - Dilution, M008 - Data Pollution	1) Pollute the data voids with wholesome content (Kittens! Babyshark!). 2) fill data voids with relevant information, e.g. increase Russian-language programming in areas subject to Russian disinformation.	Filling voids with information relevant to the topic at hand is an exercise of freedom of speech and is ethical if conducted transparently. "Polluting" or flooding the information environment is manipulative.	TA05 Microtargeting	D04
C00189	Ensure that platforms are taking down flagged accounts	M003 - Daylight	Use ongoing analysis/monitoring of "flagged" profiles. Confirm whether platforms are actively removing flagged accounts, and raise pressure via e.g. government organisations to encourage removal	The considerations here are identical to those discussed under "Deplatform Account" above. Specifically, when government gets involved in flagging content for removal, this may become coercion, depending upon the circumstance and the communications involved, and would therefore constitute a violation of the First Amendment in the U.S. See reference to Missouri vs. Biden.	TA15 - Establish Social Assets	D06
C00195	Redirect searches away from disinformation or extremist content	M002 - Diversion	Use Google AdWords to identify instances in which people search Google about particular fake-news stories or propaganda themes. Includes Monetize centrist SEO by subsidising the difference in greater clicks towards extremist content.	Manipulating search results based on an assessment that resulting content is disinformation and should therefore by suppressed is a clear impediment to the free flow of ideas. Applying monetary subsidies to favor certain results over others also interferes with the marketplace of ideas, just as subsidies on traded goods interferes with the free flow of trade in the real world.	TA07 Channel Selection	D02

				Search engines must, of course, prioritize their search results in some manner. And in some jurisdictions, such as the United States, they have broad latitude to do so as they see fit, in keeping with their publishing rights granted under the First Amendment. A review of the Wikipedia page on "Google Ads" (formerly AdWords) reveals how much power search engine companies have to decide what their users see or don't see when they search the internet: Google Ads - Wikipedia. Historically, Google has encouraged users to get a court order if they wish damaging search content removed, except in very narrow circumstances such as revenge pornography or violations of intellectual property (Google Must Remove False Information from Search Results, Says EU Highest Court | Kohrman Jackson & Krantz LLP - JDSupra). In December, 2022, however, the European Court of Justice ruled that Google must remove indexing of (or "de-reference") content that a requester can prove is "manifestly inaccurate" (Google must take down search results with "manifestly inaccurate" personal data - Lexology).		
C00197	remove suspicious accounts	M005 - Removal	Standard reporting for false profiles (identity issues). Includes detecting hijacked accounts and reallocating them - if possible, back to original owners.	See the discussion under "Deplatform Account" above. Identifying and reassigning hijacked accounts is a laudable goal but may be complex if the account was sold to an unwitting buyer. Care also needs to be exercised when removing so-called "false" profiles, since anonymity can be vitally important for political dissidents, whistleblowers, victims of sexual abuse, victims of cyberbullying. As the discussion on "Deplatform Account" indicates, this is an	TA15 - Establish Social Assets	D02

				extreme action, that should only be taken by platforms in the most extreme of circumstances.		
C00205	strong dialogue between the federal government and private sector to encourage better reporting	M007 - Metatechnique	Increase civic resilience by partnering with business community to combat grey zone threats and ensuring adequate reporting and enforcement mechanisms.	Gray-zone and hybrid threats which remain below the threshold for conventional military responses definitely require close collaboration between governments, which may have superior intelligence on the threats, and the private sector, which likely owns critical infrastructure and is in the front line of defense against such threats. Hybrid CoE Strategic Analysis 6: Countering Hybrid Threats: Role of Private Sector Increasingly Important – Shared Responsibility Needed - Hybrid CoE - The European Centre of Excellence for Countering Hybrid Threats. General awareness of the threat is an important component of societal resilience. However, any public-private partnership in the US that involves reporting must ensure that the government does not exert undue pressure on publishers in violation of the abridgement clause of the First Amendment.	TA01 Strategic Planning	D03
C00216	Use advertiser controls to stem flow of funds to bad actors	M014 - Reduce Resources	Prevent ad revenue going to disinformation domains	The digital advertising system is complex and opaque, with several intermediaries. It can be abused by fraudsters. It can also result in advertisements appearing on sites that are potentially harmful to a brand's reputation. Hence the creation of organizations such as the Check My Ads Institute and ratings services such as the Global Disinformation Index. Brands can use such services to filter out unwanted sites for their ad placements and ensure that they are not funding, for example, terrorist websites. In the U.S. for private companies such filtering is within their First	TA05 Microtargeting	D02

				Amendment rights. However, government funding of such filtering services is currently being scrutinized under the First Amendment in a lawsuit brought against the US State Department (Conservative media groups and Texas accuse US State Department of censorship | Courthouse News Service).		

Table 5 Counters which are highly problematic

disarm_id	name	metatechnique	summary	Ethical and Legal Considerations	tactic	response type
C00016	Censorship	M005 - Removal	Alter and/or block the publication/dissemination of information controlled by disinformation creators. Not recommended.	(1) Violates UNDHR Article 19 (freedom of opinion and expression); typically backfires strategically by rendering the censored content more attractive to the public. (2) In a democracy, government-imposed removal of content, or worse, removal of accounts, pages, groups, channels etc. should be reserved for clear violations of law such as child pornography or terrorist incitement to violence. The challenge here is that the law is different from one jurisdiction to the next. In the US, for example, individual freedoms are rooted in the country's historical struggle for independence from the British monarchy; the First Amendment is strictly upheld. The US Supreme Court has set the bar very high for government censorship of free speech, unless that speech is "directed to inciting or producing imminent lawless action and is likely to incite or produce such action" (Brandenburg vs. Ohio). By contrast, German law prohibits denying the Holocaust and disseminating Nazi propaganda; its laws on hate speech are deeply rooted in its own history and national identity. (3) Removing content or internet assets is even more problematic when is comes to disinformation or misinformation in which the harm to a target audience is less obvious and establishing what constitutes disinformation or misinformation is often highly subjective. In the vast majority of cases, the most appropriate response to	TA01 Strategic Planning	D02

				disinformation or misinformation is more information, not less. In the marketplace of ideas members of the public will make their own decisions about what information to trust based upon the evidence provided and/or the source of the information. Rather than removing content or assets, empowering users with a choice of transparency tools that shed light on the provenance of the information and whether automation or artificial intelligence was used in its creation or propagation is the best way to balance freedom of speech and freedom from harm. (4) Note that a distinction needs to be made between government censorship and the legitimate exercise by commercial platforms of free speech and free press rights: the US First Amendment, for example, grants platforms broad leeway to decide what types of content they want to publish or not publish - these rights are typically spelt out in clauses within a platform's Terms of Service which make it very clear that the platform can remove content or accounts at their own discretion. In practice, many platforms also enumerate certain types of "lawful but awful" content that they will not tolerate on their platform: such content may be constitutionally protected from government censorship but can still be legally removed by commercial platforms who are exercising their own First Amendment rights to freedom of speech and freedom of the press (Online CLE & MCLE | The Law of Deplatforming (talksonlaw.com)). (5) Where things get tricky in the US are situations where the government collaborates with		

				commerical platforms in the form of public-private partnerships. The danger here is that the government can use its power to coerce a commercial platform to promote or remove content on its platform. Called "jawboning", such action "is dangerous because it allows government officials to assume powers not granted to them by law" (Jawboning against Speech | Cato Institute). In a landmark case currently before the US Supreme Court, Murthy v. Missouri (formerly Missouri v. Biden), the plaintiffs argue that communication between Biden administration officials and social media platforms like Facebook and Twitter in recent years constituted a violation of the First Amendment. The defendants argue that the administration was taking responsible action to protect public health, safety and security when confronted with the challenges of a deadly pandemic and foreign attacks on elections (Appeals Court Rules White House Overstepped 1st Amendment on Social Media - The New York Times (nytimes.com)). The case will rest upon whether the plaintiffs can satisfy the "state action requirement" (state action requirement | Wex | US Law | LII / Legal Information Institute (cornell.edu)) by proving that the government's communication rose to the level of a concerted coercion campaign such that ensuing actions taken by the platform can be attributed to the state in violation of the First Amendment (Missouri v. Biden: An Opportunity to Clarify Messy First Amendment Doctrine | Knight First Amendment Institute (knightcolumbia.org)).		

				The court has agreed to hear the case in the 2023-2024 term (Murthy v. Missouri - Wikipedia).		
C00029	Create fake website to issue counter narrative and counter narrative through physical merchandise	M002 - Diversion	Create websites in disinformation voids - spaces where people are looking for known disinformation.	Inauthentic websites are a no-no as they violate the principle of transparency and presuppose deception or manipulation	TA02 Objective Planning	D03
C00036	Infiltrate the in-group to discredit leaders (divide)	M013 - Targeting	All of these would be highly affected by infiltration or false-claims of infiltration.	This is an offensive counterintelligence technique that would only be acceptable in a democracy if carried out legally by law enforcement or military personnel tasked with targeting terrorist groups, drug cartels etc.	TA15 - Establish Social Assets	D02
C00047	Honeypot with coordinated inauthentics	M008 - Data Pollution	Flood disinformation spaces with obviously fake content, to dilute core misinformation narratives in them.	Flooding the information space is manipulative. Spreading fake content is also unethical and potentially harmful.	TA15 Establish Social Assets	D05
C00052	Infiltrate platforms	M013 - Targeting	Detect and degrade	Infiltration involves covert action and is inherently deceptive and unethical. It would only be acceptable if sanctioned by law and carried out by authorized persons in situations where it was warranted due to national security, public safety etc.	TA15 Establish Social Assets	D04

C00070	Block access to disinformation resources	M005 - Removal	Resources = accounts, channels etc. Block access to platform. DDOS an attacker. TA02*: DDOS at the critical time, to deny an adversary's time-bound objective. T0008: A quick response to a proto-viral story will affect it's ability to spread and raise questions about their legitimacy. Hashtag: Against the platform, by drowning the hashtag. T0046 - Search Engine Optimisation: Sub-optimal website performance affect its search engine rank, which I interpret as "blocking access to a platform".	Any type of DDoS or flooding or SEO manipulation is anti-democratic. Such actions need to be reserved for authorized law enforcement or intelligence agency personnel with a clear remit to counter terrorists, drug cartels etc.	TA02 Objective Planning	D02
C00072	Remove non-relevant content from special interest groups - not recommended	M005 - Removal	Check special-interest groups (e.g. medical, knitting) for unrelated and misinformation-linked content, and remove it.	It is up to the special interest group itself to define its own community guidelines about what is acceptable content for the group and how to administer or enforce those guidelines. Outsiders have no business in removing content unless it is illegal.	TA06 Develop Content	D02
C00076	Prohibit images in political discourse channels	M005 - Removal	Make political discussion channels text-only.	Who enforces such a prohibition? This appears to be a major violation of the freedom of expression of politicians, advocates, lobbyists etc. Clearly, if a commercial platform wants to differentiate itself by offering text-based communication only then that is their prerogative. Whether that would be a winning market proposition is another matter.	TA06 Develop Content	D02
C00084	Modify disinformation narratives, and rebroadcast them	M002 - Diversion	Includes “poison pill recasting of message” and “steal their truths”. Many techniques involve promotion which could be manipulated. For example, online fundings or rallies could be advertised, through compromised or fake channels, as being associated with "far-	Any kind of efforts to manipulate the information environment covertly are fundamentally unethical	TA06 Develop Content	D03

			up/down/left/right" actors. "Long Game" narratives could be subjected in a similar way with negative connotations. Can also replay technique T0003.			
C00087	Make more noise than the disinformation	M009 - Dilution		This is a race to the bottom and would only serve to degrade the overall quality of the information environment further.	TA06 Develop Content	D04
C00086	Distract from noise with addictive content	M002 - Diversion	Example: Interject addictive links or contents into discussions of disinformation materials and measure a "conversion rate" of users who engage with your content and away from the social media channel's "information bubble" around the disinformation item. Use bots to amplify and upvote the addictive content.	This is social engineering which exploits cognitive vulnerabilities and automation to promote the author's content. There may be a debate to be had around authorized law enforcement personnel using such techniques in counterterrorism or counter-radicalization programs but even in these circumstances any use of psychological manipulation or covert automation in a democracy is highly problematic and should be discouraged if not banned outright.	TA06 Develop Content	D04
C00090	Fake engagement system	M002 - Diversion	Create honeypots for misinformation creators to engage with, and reduce the resources they have available for misinformation campaigns.		TA07 Channel Selection	D05
C00091	Honeypot social community	M002 - Diversion	Set honeypots, e.g. communities, in networks likely to be used for disinformation.		TA06 Develop Content	D05
C00103	Create a bot that engages / distract trolls	M002 - Diversion	This is reactive, not active measure (honeypots are active). It's a platform controlled measure.	This is an offensive counterintelligence technique that would only be acceptable in a democracy if carried out legally by law	TA07 Channel Selection	D05

				enforcement or military personnel tasked with targeting terrorist groups, drug cartels etc.		
C00106	Click-bait centrist content	M002 - Diversion	Create emotive centrist content that gets more clicks	Click-bait of any kind is a form of emotional or psychological manipulation and is not advised in a democratic public sphere.	TA06 Develop Content	D03
C00129	Use banking to cut off access	M014 - Reduce Resources	fiscal sanctions; parallel to counter terrorism	When democratic governments use their power to freeze bank accounts or otherwise sanction individuals or entities financially, this is often carried out against individuals or entities engaging in terrorist or criminal offenses. For governments to act in this way in response to problematic speech or stated beliefs, however, would in the vast majority of cases be an egregious violation of Articles 18 or 19 of the UN Declaration of Human Rights (18 - everyone has the right to freedom of thought, conscience and religion; 19 - everyone has the right to freedom of opinion and expression). Commercial banks, however, have much more leeway in most democracies to close accounts or refuse to open them without breaking the law. In most jurisdictions, the decision to offer banking or other payment account services to a customer is largely driven by a firm's commercial considerations, but also by considerations of regulatory and reputational risk. In the UK, for example, banks can choose whom they want to do business with subject to the provisions of the Equality Act: they cannot discriminate against customers based on "protected characteristics" such as "religion and belief"; but some extreme political beliefs are not protected by the Equality Act, such as Nazi beliefs. UK banks and payment	TA09 Exposure	D02

				companies are also required by UK law to provide a safe working environment for their staff, so they can legitimately refuse to do business with people who are abusive towards their staff. UK banks are also responsible for taking measures to prevent financial crime: the vast majority of UK bank account closures are either because the account is dormant or because the bank suspects financial crime. See Podcast: Debunking Debanking | Simmons & Simmons (simmons-simmons.com) and Nothing to look at here - by Frances Coppola (substack.com). In the US, there is concern that broad language about hate speech in many banks' terms of service gives staff "carte blanche authority to deny or restrict service for vague, arbitrary, or viewpoint-based reasons". See Viewpoint Diversity Score.		
C00138	Spam domestic actors with lawsuits	M014 - Reduce Resources	File multiple lawsuits against known misinformation creators and posters, to distract them from disinformation creation.	Mounting frivolous or SLAPP (Strategic Lawsuit Against Public Participation) lawsuits is an unethical practice which creates a chilling effect on public participation. If journalists or researchers are refraining from legitimate investigations for fear of frivolous lawsuits, the legal system should create a higher bar for bringing such lawsuits by passing anti-SLAPP laws. See Anti-SLAPP Laws Introduction - Reporters Committee (rcfp.org).	TA11 Persistence	D03
C00139	Weaponise youtube content matrices	M004 - Friction	God knows what this is. Keeping temporarily in case we work it out.	Assuming this is referring to YouTube content marketing matrices, or planning tools to help generate ideas for YouTube content, but it is not clear. Any type of weaponization would appear to be manipulative and unethical.	TA11 Persistence	D03

C00140	"Bomb" link shorteners with lots of calls	M008 - Data Pollution	Applies to most of the content used by exposure techniques except "T0055 - Use hashtag”. Applies to analytics	This sounds like a Denial-of-Service attack against a UR Shortening service that is used by disinformers or cyber gangs to trick people into going to an inauthentic or malicious site. It is completely inappropriate.	TA12 Measure Effectiveness	D03
C00144	Buy out troll farm employees / offer them jobs	M014 - Reduce Resources	Degrade the infrastructure. Could e.g. pay to not act for 30 days. Not recommended	This type of action would likely entail the use of economic power to lure employees of a foreign troll farm to pursue an alternative path, raising issues of interference in a sovereign state. Even in a democracy with an open employment market, attempts to hire employees away from other companies is usually done clandestinely and is therefore ethically problematic.	TA02 Objective Planning	D04
C00148	Add random links to network graphs	M008 - Data Pollution	If creators are using network analysis to determine how to attack networks, then adding random extra links to those networks might throw that analysis out enough to change attack outcomes. Unsure which DISARM techniques.	Such manipulative techniques belong to the world of offensive counterintelligence but are deceptive by nature and therefore unethical in democratic civil societies.	TA12 Measure Effectiveness	D04
C00149	Poison the monitoring & evaluation data	M008 - Data Pollution	Includes Pollute the AB-testing data feeds: Polluting A/B testing requires knowledge of MOEs and MOPs. A/B testing must be caught early when there is relatively little data available so infiltration of TAs and understanding of how content is migrated from testing to larger audiences is fundamental.	Another offensive counterintelligence technique that involves infiltrating an enemy community, studying their own techniques, and then using these against them. Does not belong in a democratic civil society.	TA12 Measure Effectiveness	D04

C00153	Take pre-emptive action against actors' infrastructure	M013 - Targeting	Align offensive cyber action with information operations and counter disinformation approaches, where appropriate.	Off limits for civilian actors in a democracy. Must be reserved for government agencies authorized to neutralize foreign enemies e.g. US Cyber Command taking action against a Russian troll farm (Cyber Command Operation Took Down Russian Troll Farm for Midterm Elections - The New York Times (nytimes.com)).	TA01 Strategic Planning	D03
C00155	Ban incident actors from funding sites	M005 - Removal	Ban misinformation creators and posters from funding sites	A ban on the funding of websites is an extreme measure that should be reserved for strictly circumscribed circumstances codified into law, such as the US ban on funding unlawful internet gambling (eCFR :: 12 CFR Part 233 -- Prohibition on Funding of Unlawful Internet Gambling (Regulation GG))	TA15 - Establish Social Assets	D02
C00162	Unravel/target the Potemkin villages	M013 - Targeting	Kremlin's narrative spin extends through constellations of "civil society" organisations, political parties, churches, and other actors. Moscow leverages think tanks, human rights groups, election observers, Eurasianist integration groups, and orthodox groups. A collection of Russian civil society organisations, such as the Federal Agency for the Commonwealth of Independent States Affairs, Compatriots Living Abroad, and International Humanitarian Cooperation, together receive at least US$100 million per year, in addition to government-organized nongovernmental organisations (NGOs), at least 150 of which are funded by Russian presidential grants totaling US$70 million per year.	Exposing front organizations used by nation-states to subvert democracies is unproblematic for civil society actors but targeting and unraveling such front organizations should be left to the national security apparatus.	TA15 Establish Social Assets	D03

C00164	compatriot policy	M013 - Targeting	protect the interests of this population and, more importantly, influence the population to support pro-Russia causes and effectively influence the politics of its neighbours	Unclear what is meant here but it sounds like covert propaganda directed at diasporas. Off limits.	TA02 Objective Planning	D03
C00202	Set data 'honeytraps'	M002 - Diversion	Set honeytraps in content likely to be accessed for disinformation.	Honey traps involve the use of deception to lure and identify targets for prosecution, to lure targets into revealing information, or to influence targets to act in a manner beneficial to the deceiver. Honeytraps are therefore highly unethical. Even in the case of law enforcement using honeytraps to track down paedophiles, the use of honeytraps is highly questionable and fraught with procedural pitfalls. See ‘Paedophile Hunters’, Criminal Procedure, and Fundamental Human Rights - Purshouse - 2020 - Journal of Law and Society - Wiley Online Library.	TA06 Develop Content	D02
C00203	Stop offering press credentials to propaganda outlets	M004 - Friction	Remove access to official press events from known misinformation actors.	In the U.S. private actors have broad latitude on deciding which press outlets they grant press passes to. The same latitude does not apply to the U.S. government, as explained in Sherrill v. Knight, 569 F.2d 124 (D.C. Cir. 1977): “arbitrary or content-based criteria for press pass issuance are prohibited under the first amendment”. As the D.C. Circuit wrote, “White House press facilities having been made publicly available as a source of information for newsmen, the protection afforded newsgathering under the first amendment guarantee of freedom of the press requires that this access not be denied arbitrarily or for less than compelling reasons.” Given that lies are largely protected under the First Amendment, for the US government to deny a press	TA15 Establish Social Assets	D03

				credential on the basis of lies is likely unconstitutional. See politico.com/f/?id=00000167-186d-def8-a56f-98ef20150002.		
C00207	Run a competing disinformation campaign - not recommended	M013 - Targeting		Fighting disinformation with disinformation is unethical in a democratic society. There may be a place for offensive information operations in wartime, but only when conducted by personnel fully authorized to conduct such operations.	TA02 Objective Planning	D07

Footnotes

[1] The military, intelligence, and law enforcement agencies of democratic national governments and international governmental organizations are granted special powers. To make ethical value judgments on the use of these powers to counter information manipulation and online harm would demand a more comprehensive analysis than this document can provide. Such an analysis would need to go beyond an International Human Rights Law frame of reference. It would need to include, for example, the Law of Armed Conflict/International Humanitarian Law, International Criminal Law, the Tallinn Manual, the Principles of Non-intervention and Sovereignty, the No-Harm Principle, and the Corfu Channel. See Tsvetelina van Benthem, Talita Dias, and Duncan B. Hollis, Information Operations under International Law, 55 *Vanderbilt Law Review* 1217 (2023), "Information Operations under International Law" by Tsvetelina van Benthem, Talita Dias et al. (vanderbilt.edu).

[2] The Declaration of Human Rights was adopted by the UN in 1948. Most of it became legally binding in 1966 with the adoption of two related treaties: the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social, and Cultural Rights. Many nations have ratified these treaties, but with exceptions. For example, the US ratified the ICCPR in 1992 subject to the reservation “That Article 20 does not authorize or require legislation or other action by the United States that would restrict the right of free speech and association protected by the Constitution and laws of the United States”. Belgium, the UK, Australia, and many other nations raised similar objections. See John Samples, *International Law and “Hate Speech” Online*, CATO Institute Blog, International Law and “Hate Speech” Online | Cato at Liberty Blog.

[3] See, for example, Thomas Healy, *The Great Dissent*, Metropolitan Books, 2013.

[4] There are several challenges to the marketplace concept in today’s media environment, created not just by the imbalance in ability to participate in the market created by large media corporations and technology companies, but also by the fragmentation and balkanization of the information environment, and the rise of generative artificial intelligence. Ideas are no longer subjected to the scrutiny of the general public but are rather propagated via polarized media ecosystems or in distinct echo chambers, such that people end up talking past each other, and the authenticity of information is often hard to ascertain. In such an environment it can no longer always be assumed that “the truth will prevail”. See, for example, Dawn C. Nunziato, *The Varieties of Counterspeech and Censorship on Social Media*, GW Law, 2021. See "The Varieties of Counterspeech and Censorship on Social Media" by Dawn C. Nunziato (gwu.edu).

[5] For analysis see Jeff Kosseff, *Liar in a Crowded Theater. Freedom of Speech in a World of Misinformation*, Johns Hopkins University Press, 2023.

[6] United States v. Alvarez, 567 U.S. 709 (2012). See United States v. Alvarez, 567 U.S. 709 | Casetext Search + Citator.

[7] New York Times v. Sullivan, 376 U.S. (1964). See New York Times Co. v. Sullivan, 376 U.S. 254 | Casetext Search + Citator.