Briefing paper

No 5, December 2023

Not One but Many Silver Bullets

Towards a Classification of Responses to Disinformation

Посолство
на Федерална република Германия
София

This publication is supported by the German Federal Foreign Office. The opinions, findings and conclusions stated herein are those of the author and do not necessarily reflect those of the German Federal Foreign Office.

Editor
Dr. Rumena Filipova
Chairperson
Institute for Global Analytics

Author
Dr. Christopher Nehring
Senior Fellow
Institute for Global Analytics

As disinformation has come to the media, social and political foreground as a phenomenon that has the power to subvert democracy, public trust and quality journalism, researchers, politicians, civil societal activists, and journalists have developed (and sometimes re-discovered) numerous responses to informational manipulation. The aim of this paper is not to suggest even more new countermeasures. Instead, it draws on the assumption that the corpus of countermeasures and analytical approaches has grown so extensive that actors who set out to build capacities or to reframe and elevate their anti-disinformation efforts are in need of a classification. In what follows, the present research offers a framework for clustering responses to disinformation into typologies that take into account thematic policy areas and the timeline for implementing countermeasures. Three outstanding challenges to the implementation of effective responses to disinformation are further charted out and include: 1) the problem of attributing the origin and source of a disinformation campaign; 2) the importance of breaking the financial streams of disinformers; 3) making optimal decisions as to whether to (always) react to a piece of disinformation.

A Typological Framework for Tackling Disinformation

A variety of stakeholders from government, civil society and international institutions can and should be engaged in the process of countering informational influence operations. While each of those types of actors focuses on their respective sphere of competence, they should also act in concert in order to implement a whole-of-government and whole-of-society approach that holistically addresses the challenges of disinformation.

The responses that can be undertaken in this regard can be grouped together and delineated according to the policy area which they refer to, including the following:

- **Legal responses**: they refer to normative, regulatory and/or legislative approaches aimed at fighting and preventing disinformation. Weak forms of legal responses start with norms, such as *codes of conduct* (e.g. the EU's Code of Conduct on Disinformation), *platform terms and conditions*, which may further lead to *platform regulation laws* (e.g. the EU's Digital Services Act) or *international sanctions* (e.g. US and EU sanctions against Russian intelligence agents, oligarchs and media). The *ban of certain content or entire media* may be regarded as the strongest legal countermeasure. A pertinent example is the ban on the broadcasting of the Russian, state-owned propaganda outlets (such as RT) in the informational spaces of EU member states after Russia's war of aggression against Ukraine.

  On the other hand, *anti-disinformation laws* making disinformation illegal have often been employed as a means by authoritarian states (such as Russia) to establish wholesale control of the media space. Liberal democracies have traditionally refrained from imposing stringent measures against disinformation due to concerns about free speech and the difficulty of providing a bullet-proof legal definition of disinformation that cannot be captured and instrumentalized by any government to oppress political opposition.

- **Security responses**: since informational operations can be carried out by foreign intelligence services, their counterparts in the Western world are involved in monitoring, analyzing, reporting and sometimes also actively fighting disinformation. Yet, in an age of large-scale digital dis- and misinformation, intelligence and other security actors are, however, only one among various actors who engage in the process of countering disinformation. As the amount of propaganda grows steadily, security agencies focus on attacks on state institutions and personnel, but not on more general activities that try to divide and weaken Western societies.

- **Technological responses**: in light of the ever-growing regulatory pressure on tech giants operating social media platforms to step up their responses to disinformation, a variety of automated solutions have been employed. These encompass *detecting and recognizing* so-called "inauthentic accounts and activities", i.e. bots and trolls, automated mes-

saging, fakes and manipulated data (images, audio, video); *marking and flagging* as well as *removing and deleting* questionable content and accounts; *decreasing the visibility of and watermarking* questionable content and accounts; *alteration* of content suggestion through *algorithms*. Moreover, the work of "*elves*" or "positive troll armies" consists of citizen volunteers pushing, sharing, liking quality content and counter-narratives against disinformation. (While this tool might be effective, it comes with certain ethical challenges as it employs and copies techniques usually used by malign actors such as automated messaging, strategic increasing and decreasing of certain content, containing a high potential for misuse).

Most automated software solutions rely on forms of artificial intelligence via machine learning or even neuronal networks or deep learning. While there is no realistic chance of countering or at least diminishing the amount of online disinformation without these measures, experience over the past years urges caution:

- ✓ Automation of recognition, flagging and removal of content is *not even nearly adequate*, as blocking genuine accounts while not omitting problematic content represents a frequent occurrence. New technologies such as *watermarking*, e.g. of AI-generated content, are likewise neither fully developed, nor fully effective.
- ✓ The overall *resources* invested by large platforms in the fight against disinformation are *insufficient.*
- ✓ More *serious and effective efforts to ban disinformation* from large social networks and platforms would *severely infringe on business models* and seriously decrease profits and revenues. This may also be the main reason why one of the most promising technological countermeasures, the alteration of content suggestion algorithms, have not been employed to a full extent yet.

- **Economic responses**: efforts to tackle disinformation have also taken place on the basis of limiting the economic ecosystem and business models behind the dissemination of propagandist messages. One such countermeasure has encompassed regulation (prohibition or ban) of advertisements running along disinformation content, including automatic placement of advertising according to the amount of online traffic. Another countermeasure is related to sanctioning individuals, organizations and businesses whose activities are (wholly or partly) geared to financing disinformation (e.g. the Wagner Group or other Russian groupings and oligarchs). Since these responses are only still imperfectly implemented, some actors, such as the Global Disinformation Index, have resorted to the publication of companies' and individuals' names who advertise on disinformation outlets. Finally, an alteration of content suggestion algorithms on social media platforms would constitute not only a content-related measure, but also automatically reduce online traffic and thus advertising revenues.

- **Public communication responses**: increasing the effectiveness of communications by state actors, but also quality media and journalism, represents a *trust-building measure* that augments societal resilience and diminishes the potency of disinformation attacks. For example, conducting strategic communications is a key remit of government structures,

involving countering malign messaging, conducting informational campaigns, national image-building and creating positive narratives. For its part, *quality journalism* has a key role to play in producing factual information and news, thus limiting the space disinformation can fill. Devoting resources, personal, time and special sections for the rebuttal of disinformation, misinformation, propaganda and conspiracy theories through *fact checking* forms another important tool in the fight against disinformation.

- **Educational responses**: specialized education in the field of *media literacy*, i.e. courses and classes on how to use and navigate the (online) information space, form a key pillar in decreasing susceptibility to disinformation as well as the risk of becoming an involuntary spreader and multiplier for dis- and misinformation. Two related approaches contributing to the process of building media literacy have included *pre-bunking*, i.e. preventive education measures about the general functioning of disinformation and concrete examples, and *inoculation*, i.e. "information vaccination" against disinformation. Both constitute efforts aimed at the use of specific knowledge and education as a pre-emptive weapon and making audiences immune or less susceptible to disinformation. As such, both must also be seen as communicative measures and activities in the field of *societal resilience building.*

Overall, it should be noted that all of these responses to disinformation can be classified as either "positive" or "negative", i.e. whether they are negatively directed against existing disinformation or positively oriented on target audiences to make them less susceptible to disinformation by providing affirmative information. Legal regulations, security measures, economic sanctions are examples of negative responses. Raising the general education of the population and instituting media literacy initiatives, on the other hand, encompass positive responses to disinformation.

A timeline of counter-disinformation responses

The types of responses to be undertaken in the process of countering disinformation can be grouped around three points in time: *before*, *during* and *after* a disinformation attack has occurred.

Countermeasures against disinformation attacks are therefore best conceptualized as a holistic process. Some countermeasures can and should be applied *before* an attack takes place based on pre-bunking, inoculation, media literacy, quality journalism, general education as well as strategic communications and resilience building efforts. While some responses, such as pre-bunking and inoculation are directed at pre-empting disinformation and are specific in their design and application, others, such as quality journalism, media literacy and general education are ongoing in their execution and are not necessarily directly connected to specific disinformation attacks. Legal measures, which are designed to prevent or at least diminish disinformation may be described as a before-measure as well as an after-measure (since usually, normative legislation and regulation tend to be a belated and generic reaction to previous disinformation attacks, but their purpose is to also prevent similar attacks in the future).

Other measures may be applied *during* an ongoing disinformation attack. These are, for example, debunking, fact-checking, but also deletions or the employment of "elves". As such, these measures react to ongoing attacks and attempt to either fight back or diminish the possible effects of disinformation. These responses may be effective as protection from singular and specific disinformation attacks, but cannot tackle disinformation as a whole or prevent similar occurrences in the future.

Finally, there are some measures that are employed in the *aftermath* of a disinformation attack. Just like legal measures, investigations, persecutions, sanctions, but also flagging and deletion measures are mostly employed after a malign informational operation has happened.

The different types of actors involved in countering disinformation may have the capacity and ability to work and employ countermeasures during all phases of a disinformation attack such as strategic communications units lodged in government structures.[1] Other actors can only contribute to specific countermeasures ongoingly or at a single point of time (e.g. in education or pre-bunking). This again underlines the necessity of a coordinated and holistic approach for countering disinformation that unites the various stakeholders tasked with countering propagandist activities.

Overall, countermeasures that are employed during a disinformation attack tend to be more specific, directed towards a concrete piece of disinformation and short-term in their effect. Measures that are used before and after attacks usually tend to counter disinformation more generally and aim for medium to long-term effects.

Towards a classification

Overall, the classification of countermeasures can *facilitate the development of a coordinated anti-disinformation strategy and policy* by grouping and modelling responses according to policy-specific issue area; the positive or negative quality of those responses; the actors that undertake them and the respective timeline for taking action (see below, on p. 7, a classification table).

Lessons learnt

A number of key **lessons** can be drawn from the classification of responses to malign informational activities:

- There is no one silver-bullet countermeasure against disinformation.
- No one actor can execute all countermeasures.
- Different actors should execute different (or the same) countermeasures at the same time.
- Despite accusations of "policing the truth" that could be levelled at governments when they conduct counter-disinformation activities, transparently operating strategic communications units on the government level (which separate their functions and work from political PR) should nevertheless exercise a leadership role and aid civil society in counter-disinformation efforts.

Table 1. Classification of measures against disinformation

Measure against disinformation	Policy area	Positive	Negative	Actor	Timeline (before, during or after disinfo attack)
Counter-intelligence	Security		x	Government	Before After
Sanctions	Legal		x	Government	After
Platform regulation	Legal Economic Technological		x	Government Tech platforms	Before During After
Regulation of algorithms and ad placements	Legal Economic Technological	x		Government Tech platforms	Before
Recognition, flagging & deletion of social media content	Technological		x	Tech platforms	After
Watermarking	Technological	x	x	Tech platforms	Before After
Elves	Technological	x	x	Civil society	During After
Strategic communications	Communication	x	x	Government	Before During After
Debunking	Communication		x	Government Media Civil society	During After
Fact-checking	Communication		x	Media	During After
Pre-bunking	Communication Education	x		Government Civil Society	Before
Inoculation	Communication Education	x		Government Civil society	Before
Resilience building	Security Communication Education	x	x	Government Media Civil society	Before During After
(Support for) quality journalism	Communication	x		Media	Before During After
General education	Education	x		Government Media Civil society	Before During After
Media literacy	Education	x		Government Media Civil society	Before

The Attribution Problem in Countering Disinformation

A key problem related to crafting effective counter-disinformation responses consists in the frequent obfuscation of who and why (actors and intent) carried out a disinformation attack or where and when it originated. The lack of confidence in identifying the source of a propagandist campaign can have direct implications for the respective countermeasures:

- Countermeasures entailing a high degree of (international) reverberation such as through sanctioning foreign actors cannot be employed without sufficient evidence. Falsely accusing actors of disinformation might discredit counter-disinformation efforts and escalate conflicts.

- At the same time, identifying actors and intent may cause significant delays in the process of crafting a response.

- Not being able to attribute a disinformation attack may weaken trust in the information space and the ability of public authorities to uphold order, enforce rules and laws and guarantee the integrity of political processes.

Similar to cybersecurity there is no easy fix for attributing disinformation. Constant monitoring of narratives, accounts and activities represents a key foundation for building up background knowledge; special technologies and software might be able trace back origins; and a fully developed strategic communications approach that also employs tools such as pre-bunking and inoculation can educate target groups and audiences about disinformation attacks, making it clear that the content of malign narratives needs to be understood and debunked even if the perpetrator cannot be identified.

Breaking the Financial Streams of the Disinformation Economy

Understanding financial incentives and business models plays a crucial role in crafting effective initiatives for countering disinformation.

- Shady PR companies make money by offering to produce and disseminate disinformation.
- Individuals can generate revenues from advertisement on their websites and social media pages.
- Large social media companies make profits from ads and the data produced by online traffic (which significantly increases with massive amounts of disinformation).
- Companies advertising on high-traffic websites, media and social media channels profit by reaching a large audience.
- Examples such as the Russian Internet Research Agency, formerly operated by Yevgeny Prigozhin, have shown that disinformation outlets are part of political corruption schemes used by kleptocratic elites.

Through all the above avenues, the economy of disinformation relies on a global market and information space. In the digital sphere and its ever-new communication technologies, information and its spread can never be fully controlled and restricted; regulation and legislation of digital spaces, such as the EU's Digital Services Act, are only valid in each national state or supranational alliances such as the EU, but never on a global scale. Actors and outlets of disinformation exploit the permeability of online information spaces, producing and disseminating disinformation in countries not affected by such restrictions and then create spillover effects. Similarly, big social media platforms operate on a global scale and in the fight against disinformation exploit regional differences between countries and regions.

Overall, restraining the economy behind disinformation may not end malign information activities, but drive up the costs for propagandist dissemination, make it less profitable while at the same time decrease the amount of digital disinformation. The challenge at hand here is that all such measures (based on advertisement and algorithmic regulation limiting automated ad placement) are in direct conflict with the basic business models of large social media networks and platforms, thus necessitating consistent regulatory pressure.

Should We Respond or Stay Silent? Communicating Anti-Disinformation Measures

A fine balancing act defines decisions as to whether to respond to and debunk a piece of disinformation or maintain silence. On the one hand, disinformation might be elevated and reach a broader audience if government institutions and politicians issue an official debunking statement or if well-known media pick up the original piece of disinformation. On the other hand, however, not responding to disinformation can create incentives and help keep costs and risks low for attackers. Although official statements and reactions might indicate otherwise, malign actors are sensitive to anti-disinformation measures. During the Cold War, Soviet intelligence, for example, monitored very attentively Western countermeasures against disinformation. As internal documents state, they felt that their measures were a) not successful anymore, b) that costs for forgeries and distribution of materials had risen, c) that they had to devote more time and work to their own disinformation while being d) constantly forced to either publicly deny or explain their information policy.[2] Furthermore, studies have also shown that effective debunking must be consistent.[3] Conversely, inconsistencies in an actor's debunking behavior can severely diminish its effects.

Hence, in the complex matrix of whether to not to debunk disinformation, a number of *ground rules* should be followed:

- A constant and careful monitoring of online activity is key to understanding whether a piece of disinformation has the potential to reach a wide audience and identifying tipping points: i.e. a debunk must be issued before an information operation becomes rife. On the other hand, attention does not need to be called out to disinformation campaigns that are not likely to gain traction.

- Debunking is more effective if it is part of coordinated communication measures, e.g. if debunking actors (on the government or civil societal level) cooperate with media and journalists.

- Official debunking by state actors carries more weight than those of media or civil society actors, but are also more likely to become subject to political attacks and accusations of censorship.

Policy Recommendations

The classification of responses to disinformation and the analysis of lingering challenges to mounting effective responses yields the following conclusions and recommendations necessary for an impactful practical application of countermeasures against disinformation:

- It is not enough to employ one or two (or five or six) countermeasures, but **all of them, right now** and **simultaneously**.
- Counter disinformation efforts need a **long-term perspective, political will, expertise** and **sufficient resources**.
- **State actors** have more resources and political weight and thus **must take the lead** in anti-disinformation measures. **State-backed strategic communications/counter-disinformation agencies**, coordination teams or task forces that work both within government structures have therefore a key role to play in spearheading initiatives that tackle propagandist activities.
- To be effective, governments need to implement a **coordinated approach** based on a clear and coherent **anti-disinformation strategy**.
- Best practice shows that a **"whole-of-society-approach"** promises to deliver the most holistic results. State actors must therefore closely cooperate with and support civil society actors and the media. This cooperation also needs a sound strategy and division of labor ("*let state actors do what state actors can do best and let civil society do what civil society can do best*").
- No counter-disinformation efforts will reach full efficiency if they do not, at least partially, **cut the economy of disinformation**.

Endnotes

[1] For more on the phenomenon and practice of strategic communications, see Filipova, R., Nehring, C., 2023, *Effective Strategic Communication for Resilient State and Society A Conceptual and Institutional Blueprint*, Institute for Global Analytics, Briefing paper (2) September 2023

[2] Nehring, C., 2019, *Kleine Brüder des KGB. Die Kooperation von DDR-Auslandsaufklärung und bulgarischer Staatssicherheit*, Berlin, p. 106

[3] Lewandowsky, S., et al., 2020, *The Debunking Handbook 2020*, Skeptical Science, p. 12.

t: +359 887 760 787
e: info@globalanalytics-bg.org
w: www.globalanalytics-bg.org
f: InstituteforGlobalAnalytics
in: www.linkedin.com/company/79841057