VIGILANT

VITAL INTELLIGENCE TO INVESTIGATE ILLEGAL DISINFORMATION

Deliverable 5.2 – Intervention Support Tool

Project Information
Project Number: 101073921
Project Title: VIGILANT: VITAL INTELLIGENCE TO INVESTIGATE ILLEGAL DISINFORMATION
Funding Scheme: HORIZON-CL3-2021-FCT-01
Project Start Date: November 1st 2022

Deliverable Information
Title: D5.2: Intervention Support Tool
Work Package: 5.2 – Intervention Support Tool
Lead Beneficiary: TNO
Due Date: 30/04/2025
Revision Number: V1.0
Authors: Neill Bo Finlayson, Elisabeth Poot, Arnout de Vries
Dissemination Level: Public
Deliverable Type: Report

Overview: The purpose of this document is to explain and showcase the Disinformation Intervention Framework and the Disinformation Intervention Framework as part of the VIGILANT platform. The Disinformation Intervention Framework is based on the work of D2.4 and D5.1 of the VIGILANT project and will also provide input for the D7.5 deliverable.

Revision History

#	Implemented by	Revision Date	Description of changes
V0.1	Elisabeth Poot	01/01/2025	Document layout and styles
V0.2	Neill Bo Finlayson, Elisabeth Poot & Arnout de Vries	14/03/2025	Finalized first draft of the document
V0.3	Kimberley Kruijver, Willem Verdaasdonk & Marcel van Berlo	31/03/2025	Internal review by TNO colleagues
V0.4	Neill Bo Finlayson, Elisabeth Poot & Arnout de Vries	01/04/2025	Processing of internal review by TNO colleagues
V0.5	Neill Bo Finlayson, Elisabeth Poot & Arnout de Vries	04/04/2025	Restructuring of the recommendations and conclusion. Adding the appendix.
V0.6	Eva Power & Peter Ivanov	18/04/2025	External review
V0.7	Neill Bo Finlayson, Elisabeth Poot & Arnout de Vries	24/04/2025	Processing of external review
V1.0	Neill Bo Finlayson, Elisabeth Poot & Arnout de Vries	25/04/2025	Final version

The VIGILANT project has received funding from the European Union's Horizon Europe Programme under Grant Agreement No. 101073921. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the VIGILANT project or the European Commission. The European Commission is not liable for any use that may be made of the information contained therein.

Approval Procedure

Version #	Deliverable Name	Approved by	Partner	Approval Date
V0.3	D5.2	Kimberley Kruijver, Marcel van Berlo & Willem Verdaasdonk,	TNO	28/03/2025
V0.6	D5.2	Eva Power	TCD	18/04/2025
V0.6	D5.2	Petar Ivanov	ONTO	18/04/2025
V1.0	D5.2	Brendan Spillane	UCD	29/04/2025

Table of Acronyms

Acronym	Definition
AI	Artificial Intelligence
C5	Context, Causes, Content, Cycle of Amplification & Consequences
D&FN	Disinformation and Fake News
DBKF	Database of Known Fakes
DISARM	Disinformation Analysis and Risk Management
DX.Y	Deliverable X.Y
EU	European Union
FIMI	Foreign Information Manipulation and Interference
GDPR	General Data Protection Regulation
HLEG	High Level Expert Group
IBRA	Indicator-based risk analysis
INT	Department D'interior – Generalitat De Catalunya
ISO	International Organisation for Standardization
IT	Information Technology
KInIT	Kempelen Institute of Intelligent Technologies
KPI	Key Performance Indicator
MO	Modus Operandi
MX	Month
NGO	Non-governmental organisation
NLP	Natural Language Processing
PA	Police Authority
SNA	Social Network Analysis
TNO	Netherlands Organisation for Applied Scientific Research
TX.Y	Target X.Y
US	United States of America
WPX.X	Work Package X.X

Table of Contents

Table of Figures

1 Executive Summary

The VIGILANT project aims to address the understanding and countering of disinformation by developing an integrated platform of advanced disinformation identification and analysis tools to cover disinformation from major sources, in all modalities and in multiple languages. It aims to meet the needs of Police Authorities (PAs) by developing an integrated platform of advanced disinformation identification and analysis tools and technologies. By using the platform, PAs can be better prepared to investigate criminal activities linked to disinformation campaigns, and ultimately even prevent them. Work Package 5: Social Drivers and Behavioural Dynamics was aimed at developing tools to support the Impact Analysis and Intervention Support for the VIGILANT platform.

This report (D5.2: Intervention Support Tool) concerns the latter goal, of offering a structured overview of potential intervention approaches that are available to PAs, to allow them to effectively counter and mitigate the effects of disinformation, to prevent it leading to harmful or criminal consequences. In order to develop the Intervention Support Tool, collaboration was of the utmost importance. The framework builds upon earlier work performed in the VIGILANT project, namely on D2.4 (the C5 model - Context, Causes, Content, Cycle of Amplification and Consequences) as its theoretical base, which was integrated into the Impact Analysis Tool of D5.1. The latter ensures that the framework will also be integrated into the overall VIGILANT platform. Further collaborative engagements were made with several Police Authorities, both inside and out of the VIGILANT consortium, and the Disinformation Analysis and Risk Management (DISARM) foundation that created a framework, which was used to compare and contrast interventions.

The framework itself consists of 29 interventions, grouped into six categories: social norms, resilience, monitoring, content moderation, deplatforming, and Police Authorities. This creates an overview of the type of intervention that can be used, and in what case or situation. In this report, all interventions are described and, if possible, academically underpinned with scientific facts. The interventions were assigned characteristics, to make filtering on appropriate interventions possible in the Impact Analysis tool. These are also described in this report.

Overall, the Intervention Support Tool leads to the conclusion that combating disinformation requires a multi-disciplinary approach. Notably, implementing (combinations of) interventions is always dependent on the specific context, which makes it near impossible to develop a fully complete and exhaustive framework. Furthermore, ethics should always be taken into consideration, to ensure that no human rights (e.g. freedom of speech) are harmed when executing interventions (e.g. such as using the ethics by design framework described in D2.1).

2 Introduction

The VIGILANT project aims to improve understanding of disinformation and how to counter it by developing an integrated platform of advanced disinformation identification and analysis tools, using state-of-the-art artificial intelligence (AI) methods. The platform covers disinformation from major online sources (e.g., social media platforms and fake ‘news’ websites), in all

modalities (text, image, video) and in multiple languages with the goal of improving Police Authorities' (PA) ability to detect and counter disinformation. This report (D5.2: Decision Support Tool) concerns the latter goal, meaning that it offers an overview of intervention approaches that are available to PAs to allow them to effectively counter and mitigate the effects of disinformation to prevent it leading to harmful or criminal consequences. In line with the objectives of work package five (WP5: Social Drivers and Behavioural Dynamics) of the VIGILANT project, the purpose of this report is to develop tools to help PAs to intervene in ongoing disinformation campaigns and inoculate susceptible groups (T5.2: Intervention Support tool), based on an understanding of the social drivers and behavioural dynamics behind disinformation campaigns (T2.4 and T5.1).

By using the VIGILANT platform, PAs can be better prepared to investigate criminal or harmful activities linked to disinformation campaigns, with the ultimate goal being prevention of criminal acts. Being equipped with reliable and accurate information about the phenomena and threat of disinformation will improve the PA's ability to safeguard communities, both online and in the physical world, against potential harmful or criminal consequences associated with disinformation. VIGILANT will also help officers to create detailed reports with relevant data on disinformation campaigns (e.g., source/spreaders, networks, targets, vulnerable groups and communities) for their superior officers, policy makers, and government security units, so that resources can be better deployed.

The combined output of WP5: Social Drivers and Behavioural Dynamics (D5.1 and D5.2) will enable PA officers to monitor and assess disinformation on a large-scale, while also providing suggestions for suitable and effective interventions. The Disinformation Intervention Framework, described in this report, is therefore closely related to other work (such as the Impact Analysis Tool, D5.2), further referred to as deliverables, in the VIGILANT project.

2.1 Overview

The work in WP5: Social Drivers and Behavioural Dynamics builds upon previous work performed in the VIGILANT project, notably D2.4, which provides a thorough theoretical base and conceptual framework of disinformation(Kruijver et al, 2023), including conceptual input for the technological tools of both WP4: Multimodal and Cross-lingual Models and Tools and WP5: Social Drivers and Behavioural Dynamics. One goal of VIGILANT is to develop a deeper understanding of the cultural and societal aspects of disinformation, in order to identify persuasive tactics and impact analysis to better understand susceptible groups. Therefore, the C5 Interaction Model (Context, Causes, Content, Cycle of Amplification and Consequences) was developed to increase the understanding of the drivers, goals, motivations, psychological dispositions and actions of these diverse actors by drawing on research from across the social sciences (see Figure 1). A key contribution of D2.4 is the focus on the interaction between different elements that influence the process of disinformation – from creation to consequences.

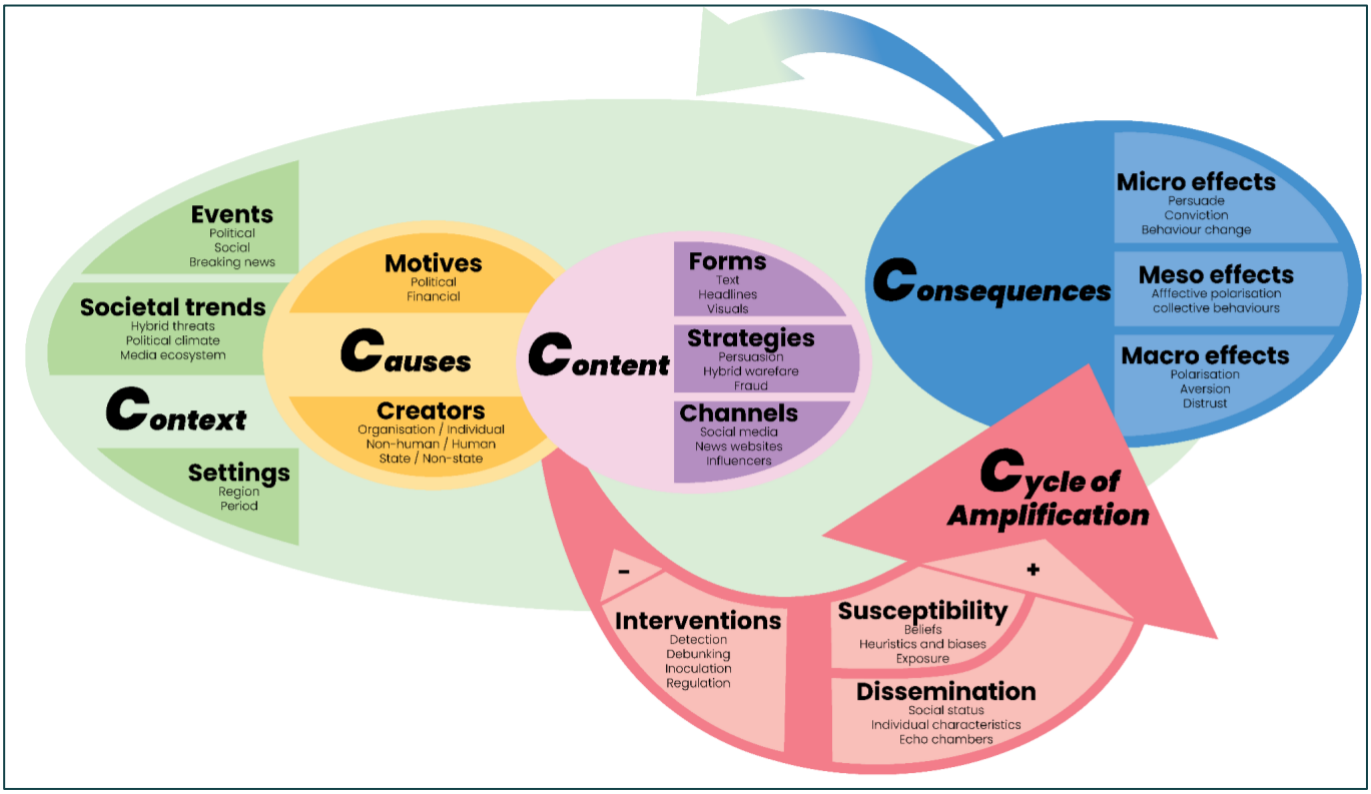

Figure 1: The C5 Model which resulted from deliverable D2.4 (Kruijver et al. 2023)

The C5 model provided the theoretical foundation for the work of WP5: Social Drivers and Behavioural Dynamics and the necessary insight to develop the Impact Analysis Tool (D5.1). The Impact Analysis Tool assists PAs in identifying, analysing, and responding to disinformation using dynamic system maps and indicator-based risk analysis. The tool applies conceptual models, which help analyse how disinformation influences individual and group behaviours, helping PAs understand the spread and impact on society and how it might lead to criminal or even terrorist behaviour (Maas et al, 2024). In close interaction with the development of the Impact Analysis Tool (see Figure 2), T5.2 developed the Intervention Support tool to assist PAs in selecting suitable interventions to counter and mitigate disinformation. This was done by the development of the Disinformation Intervention Framework, which is highlighted and explained in Section 3.2. This framework was consequently integrated into the Impact Analysis tool and thereby creating the Intervention Support Tool. The two deliverables were integrated and function as a decision support tool, the workings of which are described in the D5.1 report (Maas et al, 2024).

An overall picture of the entire VIGILANT platform is pictured below in Figure 4 and encompasses all work packages in the project. The entire platform is framed by research on social drivers and behavioural dynamics, including an ethical framework and a focus on personal data protection (see Section 2.4). The aforementioned C5 model (T2.4) is part of the social drivers and behavioural dynamics work package and provides necessary behavioural and social science background for the other work packages. The disinformation toolbox itself encompasses image and video, mapping/clustering/networking,Natural Language

Processing (NLP), interventions and impact analysis. Together, these will form the dashboard, which PAs will interact with when using the VIGILANT platform to help analyse and counter disinformation and its societal impacts.

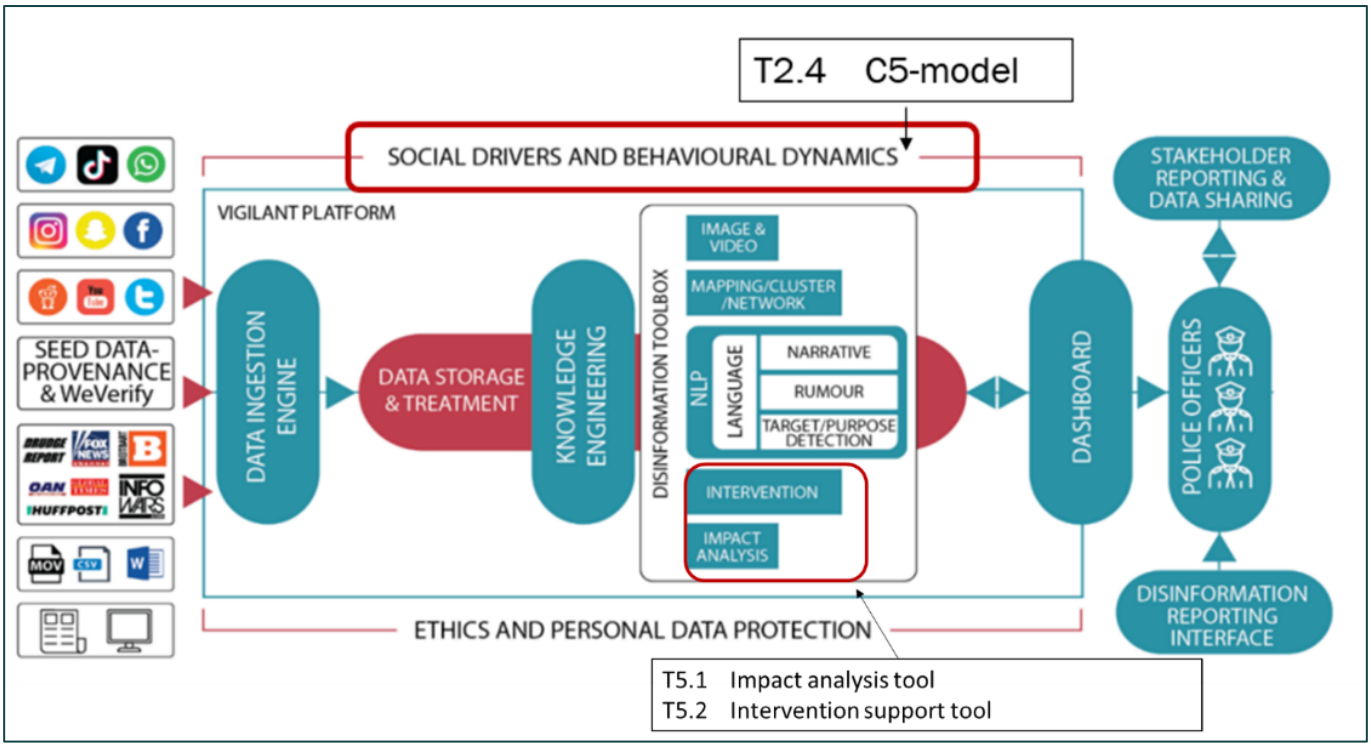

Figure 2: The entirety of the VIGILANT platform and the place of D2.4, T5.1 and T5.2.

2.2 Intervention Support Tool and Disinformation Intervention Framework

This report (D5.2) is the culmination of work carried out in T5.2. As stated previously, this comprises the Disinformation Intervention Framework which is a framework of disinformation-related interventions, and their corresponding characteristics, that PAs may employ in order to counter or mitigate disinformation and its effects. Considering that the Disinformation Intervention Framework has been integrated into the Impact Analysis Tool (IAT), the two tools (Impact Analysis Tool and the Intervention Support Tool), taken together, can be regarded as one comprehensive disinformation decision support tool (see Figures 2 and 3). This deliverable describes in detail the development of the Disinformation Intervention Framework. For a description of the integration of the framework into the IAT tool and a case study on how to use it, see Maas et al. (2024).

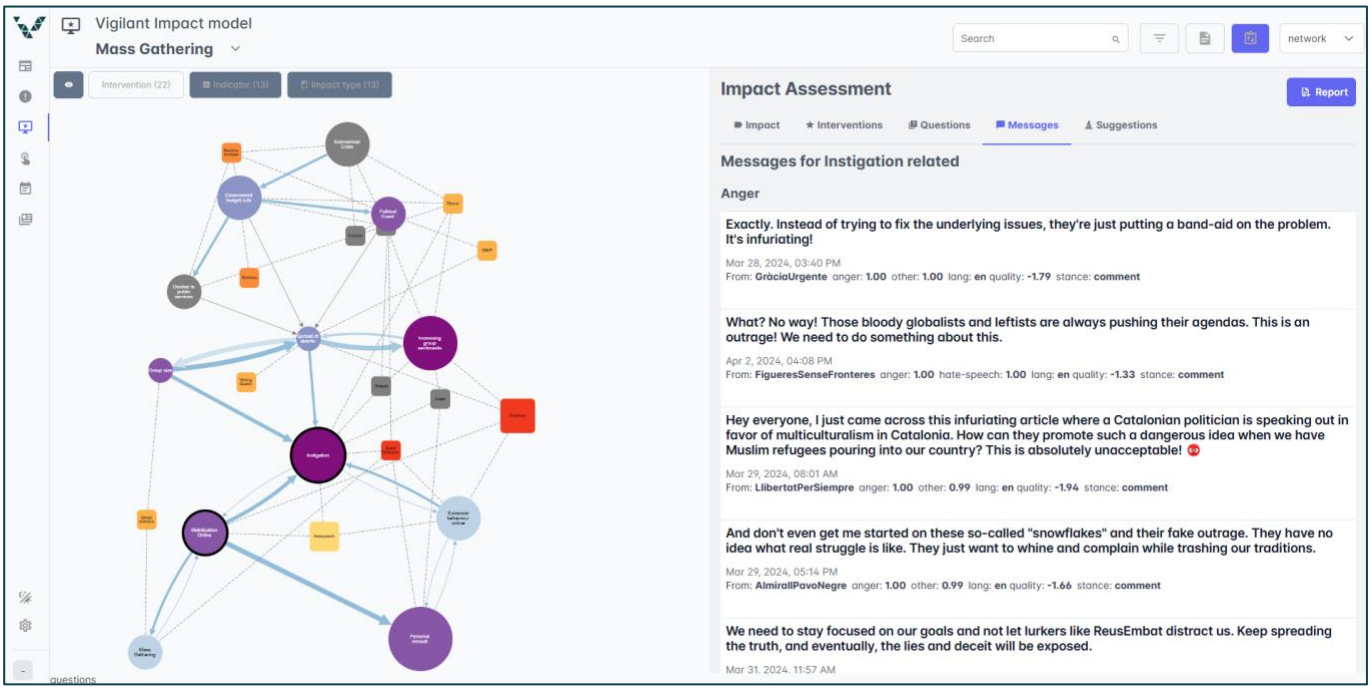

Figure 3: Impact assessment using indicator-based monitoring and system mapping for societal impacts which resulted from deliverable D5.1 (Maas et al 2024).

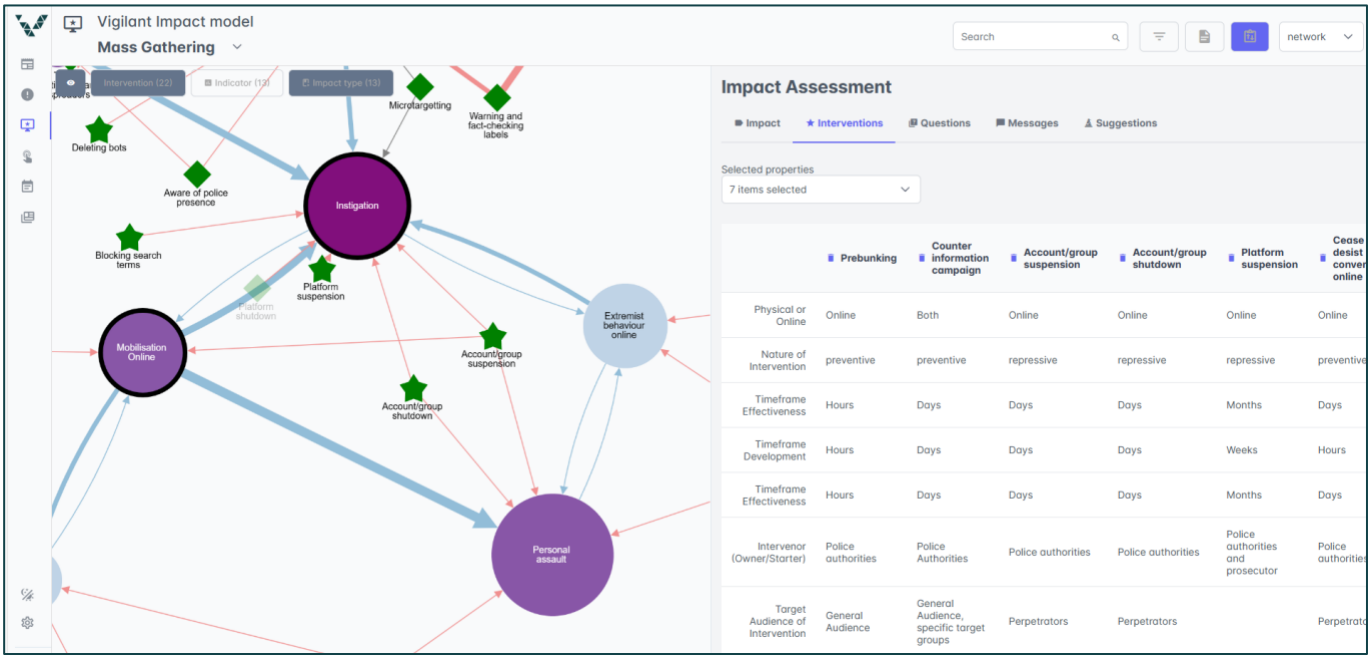

Figure 4: Decision support using system mapping and a comprehensive overview of disinformation interventions as integrated into the impact analysis tool as a result of D5.2.

2.3 VIGILANT as a Solution for Integral Policing Approach

Taken as a whole, the VIGILANT project can be seen as a solution to foster a more integral approach to policing in regard to countering disinformation. The platform and tools created as part of the VIGILANT project are designed for the whole of police, including all functions, from community police units to criminal investigations and counterterrorism. The VIGILANT supporting PAs comprises professionals from specific police units of larger police organisations in Europe, such as the intelligence units that deal with prevention of online (sexual) abuse, human trafficking, migration, corruption, cybercrime, narcotics, radicalisation, extremism and terrorism. The tasks of these units are often specific to a certain expertise, for example, those primarily working on criminal offenses and the prevention or repression of extremism or terrorism. Other police units are outward facing, focusing on communication to the public or contact with specific target groups on the streets, such as neighbourhood officers or youth police.

The different departments within police organisations, both on a regional and national level, can benefit from a common understanding (sense making) and a more integral approach towards countering disinformation. This is what the VIGILANT platform endeavours to offer PAs. Sharing insights on the different types of potential impacts and the mechanics behind them using the VIGILANT impact models (e.g., Maas et al., 2024; Kruijver et al., 2023) will increase the efforts of a police organisation as a whole, while coordinating or synchronising their efforts in both the physical and digital domain in both a repressive and preventive fashion. In addition to the PAs that are members of the VIGILANT consortium, VIGILANT has established a Community of Early Adopters (CoEA). This community comprises European PAs who have an interest in following the project and potentially adopting the VIGILANT platform at the end of the project. The engagement of the CoEA, through online workshops, has allowed us to obtain valuable feedback from a large variety of PAs throughout Europe.

2.4 Ethics

VIGILANT holds ethics and privacy concerns at its core , ensuring that relevant rules are enforced during the project's lifetime. In the field of disinformation, it is particularly important to be aware of ethics and privacy concerns at all times and apply both by conforming to existing EU law, standards and guidelines. More details on the ethical considerations for the interventions as part of the VIGILANT platform as a whole are described in D2.1 (Ethics Framework), D2.2. (Ethics guidelines for Police Authorities) and D2.3 (Ethical Oversight Report).

Ethical concerns and implications relating to countering disinformation for PAs were considered throughout the development of the Intervention Support Tool and criteria that incorporated ethical considerations were integrated into the framework itself. When listing the interventions and assigning the characteristics, 'Considerations' (see Section 5.4) were created in order to highlight interventions that might cause questions around the ethics of an intervention. Thus, interventions that are close to censorship, or may cause a backlash from the public, are flagged.

2.5 Reading Guide

The outline of the report is as follows. In Section 3, the research method and the development of the Intervention Support Tool is explained. In Section 4, an overview of all the interventions in the Intervention Support Tool is provided, with corresponding explanations and reasoning and, if relevant, scientific background. In Section 5, the characteristics of the interventions are explained. In Section 6, conclusions are drawn from the work that was carried out on the interventions, and considerations, limitations and recommendations are discussed.

3 Methodology

The Disinformation Intervention Framework was developed to be integrated into the Impact Analysis Tool of D5.1. This section focuses on the methodology used to develop the framework, the collaborative structure of the VIGILANT project and the main findings from other work packages that supported the development of D5.2. This section firstly sets out a definition of disinformation before outlining the overarching purpose of the VIGILANT project, then the extent of collaboration in the project before describing the development of the Disinformation Intervention Framework in the final three sections.

3.1 Definitions

As stated, the goal of VIGILANT is to aid PAs in dealing with disinformation. However, one of the problems of disinformation is the difficulty of defining it. The VIGILANT project adopts the definition set forth by the European Commission:

> Disinformation is false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit. Misinformation is defined as false or misleading content shared without harmful intent though the effects can still beh armful (Commission, 2018).

Disinformation and misinformation are often used interchangeably. However, as the above quotation shows, this is inaccurate: the intention to cause harm is an inherent feature of disinformation and sets it apart from misinformation. However, when it comes to countering the effects of disinformation, the distinction between mis- and disinformation is perhaps of less importance. That is, when countering the effects of disinformation, the intention of the sender (whether it was harmful intent or not) is of less concern than understanding whether the nature, route and effect of the message (the false information) is effective in stimulating attitudinal or behavioural change. As such, for the purposes of this paper, research was carried out on the assumption that any countermeasures or interventions that were found to be effective against misinformation will hold against disinformation as well.

3.2 Development of VIGILANT'S Intervention Support Tool

To start the development of the Intervention Support Tool, a framework was developed. This resulted in the Disinformation Intervention Framework, which was developed using existing literature of multi-disciplinary research on the phenomena of disinformation, supplemented by interviews with PAs and experts working in the field of disinformation. In terms of literature, a foundational source for developing the Disinformation Intervention Framework was the DISARM (Disinformation Analysis and Risk Management) framework. This framework was developed by the DISARM foundation, a not-for-profit organisation located in the United States and also working across Europe with several partner organisations (DISARM Foundation, 2024). DISARM provides a framework to better identify and respond to malign information influence operations, helping defenders to analyse and share information about the nature and scale of information threats. DISARM comprises two frameworks, the

Red Framework and the Blue Framework, which provide a common language for documenting influence operations. DISARM Red describes incident creator behaviours and DISARM Blue describes potential response behaviours.

The DISARM frameworks were developed in such a way that it splits general influencing strategies into micro tactics, which aligns with the structure of the present Disinformation Intervention Framework. Therefore, the DISAM frameworks were analysed to determine if any interventions contained in either the Red or Blue framework would be applicable for this present Disinformation Intervention Framework. As a result, some interventions from the Blue DISARM framework were implemented into the present VIGILANT intervention support tool and have been referenced as such throughout this report. The interventions in the Blue DISARM framework are rather specific, whereas the interventions in the framework are more broadly clustered. Some of the Blue DISARM interventions might therefore be suited for the framework but are not included due to the difference in scope. The Red DISARM framework mostly concerned adversarial actions (those carried out by adversaries with malicious intent to cause harm) which were not relevant for D5.2. For more details on which interventions derive from DISARM or from other sources, refer to the supplementary materials in Appendix A.

Another important touchstone in the literature for this review, was a toolbox of individual-level interventions to counter misinformation (see Section 3.1 on relevance on misinformation interventions) (Kozyreva et al., 2024). Although this was an informative resource which helped inspire the structure of the Disinformation Intervention Framework, there is far less direct overlap than with the DISARM framework. Full details of the interventions and their sources is available in Section 4 and 5 below, as well as the framework interventions table in the supplementary materials (link to the excel doc)

Aside from using existing literature, the Disinformation Interventions Framework was developed and refined through conversations and several workshops and interviews with six PAs from six countries, both inside and outside the VIGILANT Consortium (the Community of Early Adopters). During two physical workshops and two online workshops and four extra bilateral interviews, the authors asked participants about their concerns about disinformation, what countermeasures they were already working with and the extent to which they are developing new measures. During these workshops, the authors provided a list of the interventions based on the literature search, as described above, and asked them to note or mark any additions on current or past interventions they use or used.

As part of the approach to WP5: Social Drivers and Behavioural Dynamics, several workshops with PA Officers and subject-matter experts (researchers and experts from within the consortium) were held in order to ensure that the intervention support tool would aid PA officers in their decision-making process to decide on effective interventions in specific cases (called ‘knots’ in the VIGILANT tool). The online and offline workshops and bilateral interviews mentioned also highlighted several pressing issues associated with disinformation detection and analysis that go beyond the PA Officer’s own decision-making process. For instance, the lack of collaborations within their own police organisation or with other partners (in their country or internationally), such as national (cyber) safety centres, to counter disinformation.

Once the list of interventions was finalised (see section 4 for an overview of all interventions in the framework), characteristics were attributed to the interventions (see section 5). These characteristics act as an index to specify the nature and conditions of each intervention and make them more applicable for PA Officers. This means that, in the framework, interventions can be filtered and sorted according to the context and need of the Police Authorities.

The Disinformation Intervention Framework was integrated into the Impact Analysis tool of D5.1, meaning that all of the interventions contained in the framework are available in the tool for assessment and analysis as potential ways to counter disinformation. This is what forms the Intervention Support Tool (D5.2) This enables analysts to map out potential effects of various interventions and their inter-relationships (Maas et al, 2024). A more detailed description of the impact analysis tool can be found in Maas et al. (2024), D5.1 of the VIGILANT project, which outlines the functionality of the tool and role of the user.

In Figure 5, the top category of the disinformation intervention framework is pictured. The framework is structured similarly for all the categories, which are detailed in Section 4. The entire framework is available at request.

Intervention			Type			Timeframe			Execution						Considerations		
			Physical or online	Preventive and/or reactive	Escalation level	Timeframe development	Timeframe execution	Timeframe effectiveness	Intervener (owner) KAPs	Mandatory participation (pa...	Possible intervention collabo...	Target audience of intervent...	Targeted outcome of intervent...	Scalability of intervention	Judicial framework or pub...	Risks of intervention	
Social norms	Social norms	Leveraging social information to encourage people not to believe, endorse, or share misinformation. People mirror (consciously or unconsciously) new another's attitudes.	Both	Preventive	E1	Years	Years	Permanent	Police authorities	Government	NGOs, Service providers, Education	General audience	Engage others in disinfo intervention	Low	National constitutional law (freedom of speech)	Distrust in the police/government, Polarisation in society	Context
	Stimulating social cohesion online	By creating the feel of a society online, people will mirror social behaviour that they show in real life, so call each other out on offensive language and feel a certain social pressure to behave	Online	Preventive	E1	Months, Years	Months	Permanent	Police authorities	None	Government, Education, NGOs	General audiences	Normative shift	Low	National constitutional law (freedom of speech)	Polarisation in society, Distrust in the police/government, Reputation damage	Context
	Counterspeech	Counterspeech is countering offensive speech by calling out the speaker.	Both	Preventive	E1	Weeks, Days	Days	Permanent	Police authorities	None	Government	General audiences	Normative shift	Low	National constitutional law (freedom of speech)	Distrust in the police/government, Reputation damage, Polarisation in society	Context
	Stimulating online community roles	By creating online community roles, you can create the role of a community moderator, so people will feel	Online	Preventive	E2	Months	Weeks	Permanent	Police authorities	None	NGOs, Education, Service providers	General audiences	Normative shift	Low	National constitutional law (freedom of speech)	Distrust in the police/government, Reputation damage	Context

Figure 5: A screenshot of the first category of the disinformation intervention framework.

4 Disinformation Intervention Framework

Based on literature research and input from PAs, an initial list of interventions was started. During the research, the list was extended, and characteristics were assigned to each intervention. As previously mentioned, the interventions were sorted into six categories: social norms, resilience, monitoring, content moderation, deplatforming, and PAs. This was done to create a better overview of the type of intervention that can be used and in what case or situation. Each category and the interventions it contains are described in the following sub-sections.

4.1 Category 1: Social Norms

One category of interventions that can be used to combat disinformation relates to the adjustment of social norms. By leaning on social norms through what is called a social norming approach, it is possible to correct falsehoods, misconceptions or conspiracy belief norms using normative feedback (e.g. Cookson, Jolley, Dempsey & Povey, 2021). In essence, one can leverage social information to encourage people not to believe, endorse or share disinformation because people mirror the attitudes people, perceptions and actions of those around them. This also applies to disinformation in the online world whereby the communication of positive normative (socially desirable) information through a social norming approach can help in combatting the effects of disinformation (Ecker et al., 2023; Gimpel, Heger, Olenberger & Utz, 2021). This is a high intensity effort but can yield long-term results. Socially normative interventions include stimulating social cohesion online, counterspeech and stimulating socially acceptable online community roles. It is important to note that leveraging social information can potentially result in ethical malpractice, due to censorship or privacy violations.

4.1.1 Stimulating Social Cohesion Online

By nurturing a feeling of society in their online activity, people begin to mirror social behaviour that they show in real life. This means it may be possible to nurture an online environment that mirrors normal social pressures whereby, for example, people may call each other out on using offensive language or behaving inappropriately. This stimulation of social cohesion online can occur through three main channels: networks, information, and norms (González-Bailón & Lelkes, 2023). Indeed, social media can promote “positive cross-cultural and intergroup interaction, increase belongingness, and facilitate self-expression” (Selim and Popovac, 2024).

4.1.2 Counterspeech

Counterspeech is countering offensive speech by calling out the speaker and urging them towards a more socially acceptable form of speech. Counterspeech is effective against different types of unwanted speech and seeks to counteract potential harm that is brought about by the unwanted speech (Cepollaro, Lepourte & Simpson, 2023). By correcting unwanted speech, it creates a new social norm, in which the harmful speech is reduced. It was found that by introducing counterspeech to Twitter conversations, a more balanced public discourse emerged which helped quell hateful rhetoric in online discourses on German Twitter, for example (Garland et al., 2022).

4.1.3 Stimulating Online Community Roles

Similarly, by stimulating online community roles people are encouraged to take social responsibility in their online activity. People will therefore feel more protective towards the communities they form and more likely to protect them. Research suggests that fostering a sense of community on social media can help to counteract the spiral of silence (in which the willingness of opinion sharing is determined by perceived public opinion) by facilitating open expression and sharing of diverse opinions and voices (Laor, 2023). The live-streaming platform Twitch, where servers are monitored by a moderator who is responsible for adhering to the community guidelines, provides a good example of how such communities can be fostered and managed (Seering & Kairam, 2023). A study has found that increasing user engagement in content moderation has a positive effect on decreasing harmful content (Wang & Fu, 2023). An intervention could therefore be encouraging more person-based moderation. It is, however, important to note that this moderation is still dependent on the person that is assigned to moderate. If this person is ill-intentioned, or does not take their role seriously, there may be no effect, or indeed an adverse effect.

4.2 Category 2: Resilience

To reduce the impact of disinformation, it is important to strengthen the resilience of potential targets. In other words, strengthening society's ability to resist and counter disinformation reduces vulnerability to the threat. PAs can play a role in this societal resilience-building by coordinating awareness campaigns, cooperating with communities and collaborating with education partners. Such interventions can provide long-term sustainable solutions, however, they can also be very costly in both labour and time. The aim of resilience-based interventions is mostly based on improving general media literacy.

4.2.1 Counter Information Campaign

Refuting and rebuffing through counter information campaigns can help create alternative narratives, based on truth, that challenges people's beliefs in the harmful one. Research shows that truthful communications that engage people with narratives on a psychological level are more effective than mere fact-stating (Bateman & Jackson, 2024). By tapping into deeper emotions, core values and humans' need for social inclusion, counter campaigns can make misinformation unappealing and, if implemented well, can penetrate even the most extreme audiences (e.g. Davey, Tuck & Amarasingam, 2019). However, to be effective, this tactic of counter information campaigns requires elements of fact-checking, debunking and audience analysis (Bateman & Jackson, 2024). Importantly, it also requires having a disinformation countermeasure plan. Authorities should endeavour to create a policy and execution plan for disinformation response, before it's needed, including connections and contact details to speed up this process, as well as estimations of the expected counter action.

4.2.2 Prebunking / Inoculation

Rather than countering the occurrence of mis- and or disinformation, another way to build resilience is to make people less susceptible. In other words, inoculate them against misinformation. Inoculation (also referred to as prebunking) is a preventive measure. While debunking relies on proving certain misinformation 'facts' are not true, prebunking aims to prevent false information from taking root in people's minds in the first place (Roozenbeek & van der Linden, 2019). One method that builds resilience is through improving media and information literacy among citizens, particularly the promotion of good news habits and critical thinking (Nygren & Ecker, 2024). Training and education through gamification and simulations have proved effective (Roozenbeek & van der Linden, 2019; Nygren & Ecker, 2024). Prebunking, inoculation and media literacy are grouped in this report for efficiency but are often considered as distinct interventions (McBride et al., 2024).

4.2.3 Debunking / Verification and Fact-Checking

Going further than merely detecting misinformation, debunking requires not only identification but also the correction or rebuttal of falsehoods. It is often referred to as fact-checking (McBride et al., 2024). Research shows that introducing corrective information as a form of rebuttal against misinformation is far more effective in reducing the likelihood of people to believe it than merely labelling the article as false (Chan, Jones, Hall Jamieson & Albarracin, 2017). This process can be time and labour intensive because verifying the validity of information is difficult and the life cycle of information online is fast. Furthermore, the effects of debunking vary depending on numerous factors, such as the detail of the rebuttal, the time elapsed before rebuttal and the reasoning behind people's beliefs in the misinformation (Ecker et al., 2022).

4.3 Category 3: Monitoring

Although the monitoring of online content and behaviour may seem to be a passive act, it is not an intervention in itself, as the actions undertaken as part of the monitoring can provide insight into the workings of disinformation and can have a deterrent effect. For example, if monitoring by Police Authorities as a police practice is communicated to the wider public or is public knowledge, it can have a deterrent effect to start with, since people conscious of online surveillance or monitoring are likely to behave more appropriately than in unsupervised environments (Stoycheff, Liu, Xu & Wibowo, 2019). Monitoring can be assisted by technology (e.g. Hunt, Agarwal & Zhuang, 2022), but can also be conducted by human analysts individually, perhaps oriented around themes or certain organisational structures. Furthermore, spotting or locating and identifying criminals online in itself can be seen as an intervention as part of common practice in law enforcement activities. Here too, it is important to note that PAs need to adhere to the their national laws, and subsequently international laws, on monitoring and intelligence gathering, as unlawful monitoring or tracking can have serious consequences for basic human rights (see also the ethics considerations in 2.4).

4.3.1 Social Media Content Analysis

Content analysis for social media is the process of tracing and tracking disinformation online, either strategically (e.g. monitoring of media and societal trends or trends on criminality), tactically (e.g. monitoring of social or criminal

phenomena/movements) or operationally (e.g. monitoring of groups or individual people) in order to analyse the social media content in question (e.g. Lai & To, 2015). Current practices of social media content analysis can be done by specific police units, such as terrorist or crime units, but also by a communications department, operational centres (such as real-time crime centres that monitor online activity for crisis situations) or neighbourhood officers with some form of online presence (OSINT Industries Team, 2024). This monitoring can be done by analysts, media watchers or individual officers using their phone or PC as part of their specialist tasks or routine tasks such as surveillance, incident handling and crowd control, including rumour control. It can also be conducted to analyse online criminal phenomena or an individual criminal investigative case. Analyse of this kind can, for example, comprise narrative analysis, aspect sentiment analysis or social network analysis conducted using semi-automated analytical tools and software.

4.3.2 Social Network Analysis

Social network analysis (often named SNA) is a common police practice that involves analysing new criminal phenomena or conducting police intelligence and/or criminal investigations. It is performed to understand the network of people, groups or organisations and make sense of how crime scripts or modus operandi (MO's) in criminal acts come about and are organised (Hollywood et al., 2018). This provides additional insight into where in the network, and who to target when developing intervention strategies. For most PAs, this would be a new strategy of linking SNA to disinformation, but the tool described in D5.1 on the Impact Analysis Tool (Maas et al., 2024) does include these types of analyses.

4.3.3 Infiltration of Nefarious Groups

Going 'undercover', perhaps using virtual agents online as a means of automating this effort, is a common practice in several European PAs. It can either take the form of a police officer creating a fake account and joining a group, or an automated virtual agent joining and reporting a group (e.g. Evans, 2023; In Cyber News, 2024). This is not a method that is used lightly and is one of law enforcements most serious undercover power, as it is a highly costly and highly invasive operation (Kruisbergen, 2021). Therefore, in most cases, approval from public prosecutors is needed to use this type of intervention. The method often involves the use of a social media account through an alias to blend in with members of a group or forum with the aim of infiltrating specific target groups. There are differences between passive infiltration - where no action besides looking in are taken - and infiltration where active communication with other members of the group or a wider audience is established. This depends on the context of the investigation and is subject to very strict legal and ethical guidelines used by PAs. For example, provocation or illegal acts (such as hacking an account or sharing illegal content to blend in) is either not allowed at all or is subject to a very strict oversight regime from public prosecution or other legislators.

4.3.4 Tracing Instigators and Spreaders

The use of social media monitoring and analysis to find the instigators of disinformation or hate speech campaigns is about getting to the source of disinformation. In other words, tracing and identifying the largest spreaders or actors orchestrating the campaigns behind the scenes. This may entail investigating, locating and seizing botnet servers through the proper legal

procedures, such as notice and takedown procedures. Social network analysis can help in performing this task (see above), but often this is not enough as accounts use fake names or are anonymous. Hence, other interventions such as tracing IP addresses or other means to attribute real world identities to these accounts is needed as well as approval of a public prosecutor (Baraz & Montasari, 2023). However, for some international platform's international requests from PAs for legal assistance are not granted. Either the platform is situated in countries where international legal collaboration is not present or it is, but platforms can deny these legal requests if legal criteria are not met to protect the rights and privacy of customers.

4.4 Category 4: Content Moderation

The moderation of online content is a versatile method of intervening in disinformation. Content moderation consists of governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse (Grimmelmann, 2015). In the context of disinformation monitoring, content moderation entails tracking or watching content on online platforms, to ensure that the content is not (potentially) harmful, and if content is considered problematic, to take measures to mitigate the potential effects. This can consist of e.g. adding warning labels to certain content, but also to removing content altogether. There is difficulty in measuring the effectiveness of content moderation as an intervention against disinformation. This is because of limited collaboration between the service providers and both analysts and scientists, because of various definitional, methodological and stewardship challenges (Yasmin Green, 2023). Stewardship challenges are on an individual level and include difficulties in managing resources, such as time. Those who monitor disinformation often have to deal with an enormous amount of messages and manual moderation is nearly impossible. The following interventions are based on content moderation and the potential subsequent action of the moderation.

4.4.1 Warning and Fact-Checking Labels

The existence of 'credibility badges', or warning and fact-checking labels that can accompany social media posts can be effective interventions for counteracting online misinformation (Prike, 2024). They show the users of social media that the content they are interacting with may contain false or otherwise harmful information. Users are less likely to share posts when they know content is not reliable. The use of labels works especially well in combination with other interventions, such as social norms and friction. These labels could include fact checked information, but since that is difficult to implement on a large scale, warnings about content that aims to slow down the sharing process should suffice. An example of this is the community notes on X, where users can indicate whether information is questionable. The notes are, however, rarely visible to users of the platform (Nyariki, 2025).

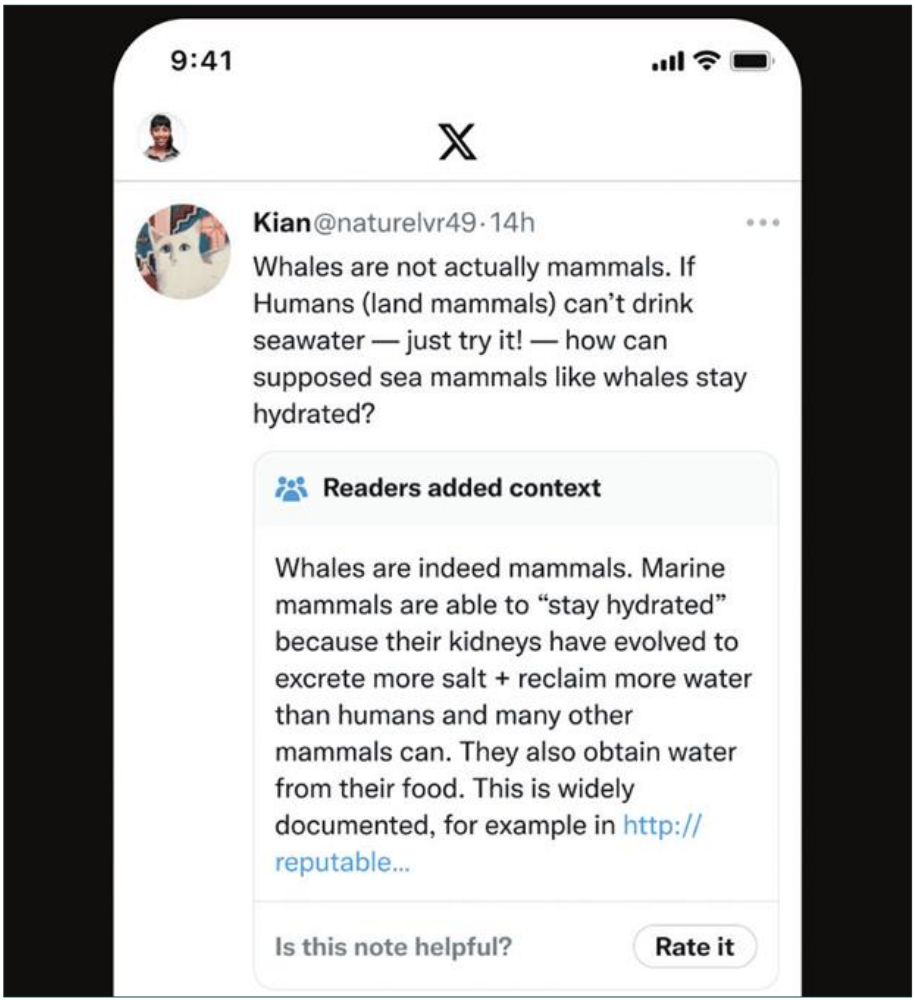

Figure 6: Fabricated example of a community note on X. Source: Community notes.

4.4.2 Microtargeting

Microtargeting is the use of online personal data in order to target individuals and show them highly personalised messages, especially in a political context (Almog Simchon, 2024). This could theoretically be reversed, to show perpetrators of disinformation targeted messages to warn them away, or to target intended audiences to make them more resilient against disinformation. This can be considered highly unethical, as it alters the right to free speech, the transparency of the internet and can be considered as censorship.

4.4.3 Blocking Search Terms

Service providers can have the option to block certain information. Google Search can, for example, block search terms that violate the content guidelines of Google to protect its users against harmful content (Google Services, 2023). A recent example is the update to the Search Algorithm to tackle AI-deepfakes, by altering the algorithm to filter out AI-generated sexually explicit

content (Irwing, 2024). Blocking search terms when incidents occur can limit the availability of information, thus limiting the ability to spread the information.

4.4.4 Redirect or Divert

By redirecting and/or diverting people away from original posts, it is possible to deter the sharing of false information, which is also referred to as creating friction as it makes the process of sharing information online less smooth. The use of data friction can alter a person's actions online (Bates, 2017). Altering algorithms is a method that can be useful for this but requires collaboration with service providers who could be less inclined to do so.

4.4.5 Content Blurring

Content blurring involved modifying content in order to hide certain content by blurring words or blurring pictures. This could be to protect the privacy of people in these texts, pictures or videos. This is less invasive than deleting content out right, as it does not restrict access, but might limit the amount of new viewers of the content and protect people's rights or prevent further damage. The content is still accessible, but it requires an active interaction in order to see the content, which might limit the amount of viewers. A fabricated example from Meta can be seen below in Figure 7, as part of the service provider's efforts to limit potential false information on Instagram (Meta, 2019).

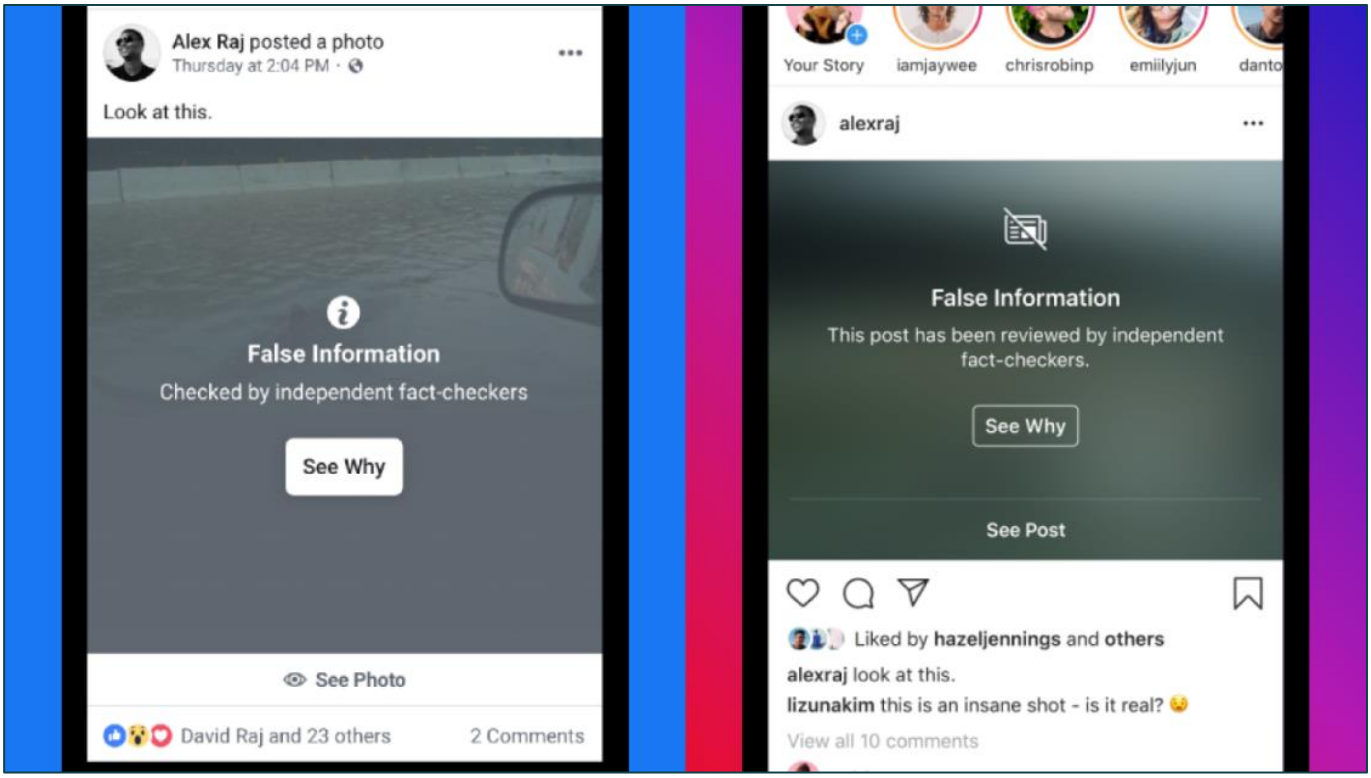

Figure 7: A fabricated example of an Instagram post is modified by adding a blur and a warning to the content. Source: Bell, 2023.

4.4.6 Content Removal

Content removal involves the (permanent or temporary) removal of content in order to limit accessibility to content that is labelled as disinformation or potentially dangerous. This restricts access to information completely and therefore goes hand in hand with censorship concerns. This intervention therefore needs to be exercised with caution. It also depends on the service provider or the social media platform in question, who can actually remove the content. There are also different pathways for removing or requesting to remove content. Often, service providers have terms of service that a user agrees with, and if content does not comply with this, it will be removed.

In addition to from the service providers themselves, PAs and other legal entities, including citizens, can also legally request to remove content on most online platforms. On larger platforms, special procedures exist for public authorities. These platforms also often choose to be transparent by publishing transparency reports on the removal of content or accounts, including actions taken at the request of governments and, specifically, law enforcement agencies — as seen, for example, in the Google Transparency Report (see Figure 8 below).

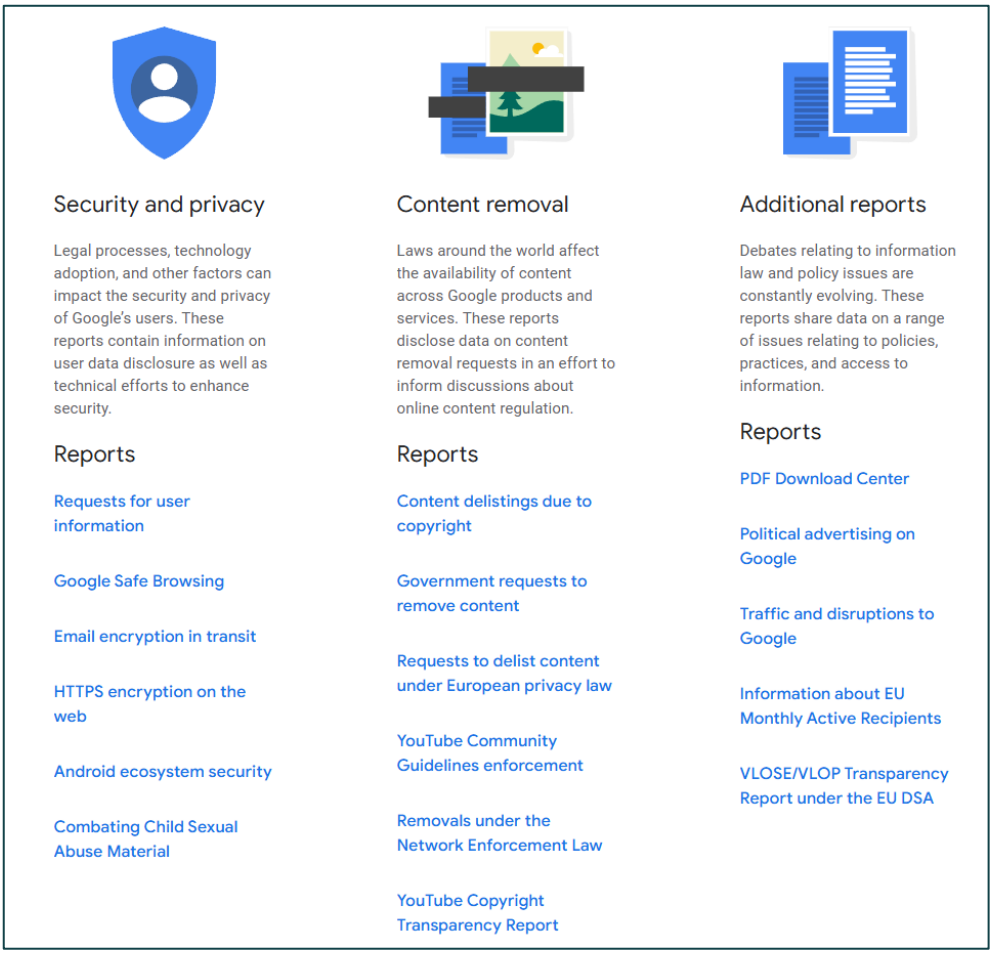

Figure 8: Example of a transparency report for measures taken in online platforms, in this case from Google.

4.4.7 Disabling Platform Features

Blocking certain features on a platform to stop harmful content from spreading is another online intervention. The messaging platform WhatsApp, has for example, limited the amount of people you can forward a message to in order to limit the spread of misinformation through mass messages (Hern, 2020). Other examples include (temporary) switching of features to earn money with the platform or map messages on a map. For example, Apple and Google temporarily disabled live traffic data in October 2023 to ensure the safety of communities (Business and Human Rights Resources, 2023). They did so to make it harder to track where large amounts of people were gathering so those locations could not be used for targeted attacks.

4.4.8 Dilution

Dilution is an intervention that involves PAs creating their own content in order to dilute disinformation. This can be done through buying advertisement space or countering with own generated content or hijacking a hashtag. Overwhelming a hashtag can work on multiple social media platforms and has been done by people who protest the message of the original hashtag. An example is the hashtag #WhiteLivesMatter that gained popularity after the killing of George Floyd, which sparked Black Lives Matter protests all over the world, to oppose the Black Lives Matter movement. In protest against the White Lives Matter hashtag, K-Pop (Korean Pop) fans flooded the hashtag with music and memes. This made it harder for #WhiteLivesMatter to gain traction (Gordon, 2020). Dilution aims to make malevolent content harder to find, to reshape the narrative around the hashtag and to de-amplify the message, as some algorithms may deprioritise the hashtag if the content appears to be less engaging (Agency, 2022).

4.5 Category 5: Deplatforming

Actors online that post, share and engage in extreme, dangerous or harmful content - including disinformation – or service providers that facilitate this, can be deplatformed. In general, this is the act of disabling or removing a social media account to make the content inaccessible (Rogers, 2020). However, the specific act of deplatforming can take many forms, as the following section outlines. The act of deplatforming has sparked intense debate in Europe and the US over the use of such measures through a perceived liberal bias of online platforms (Rogers, 2020), the potential infringement on the right to freedom of expression (Brommell, 2021) and the side-effects of pushing nefarious actors underground online (Bryanov, Vasina, Pankova & Pakholkov, 2021). These are important considerations for the role of law enforcement in deplatforming (Bromell, 2021). Furthermore, the ability for law enforcement to deplatform actors online is dependent on the cooperation and action of the online platforms themselves. The effectiveness, next to the ethics, is also under debate, as the removal of content may lead it to pop up on other platforms. The following interventions are specific actions that can achieve the broader goal of deplatforming.

4.5.1 Account / Group Shutdown

Account/group shutdown involved the shutting down of an account or group that is spreading disinformation. This often is a legal request from Police Authorities to online platforms or service providers, but requests can also be done by other government bodies or even citizens. Procedures are similar to (temporarily) removing content (see 4.4.6), but the criteria for removing accounts, group accounts or channels, are stricter and could have more impact upon individuals. An example of this is the shutdown of the QAnon movement across Facebook and Twitter (Allyn, 2020; Frenkel, 2020). The QAnon movement is fraught with disinformation and has led to multiple violent outbursts in both the U.S. and Europe. By shutting it down, an attempt was made to mitigate the harmful effects of this group.

4.5.2 Account / Group Suspension

Another intervention is suspending an account or group that is spreading disinformation. This often starts as a legal request from Police Authorities to remove an account or even a group, where online platforms or service providers decide to suspend the account until a final decision can be made or to simply use this method to suspend it for a period of time. Procedures are similar to (temporarily) removing content (see 4.4.6), but the criteria for suspending accounts, group accounts or channels are stricter and could have more impact upon individuals. An example of an account suspension is the suspension of the Trump campaign account because of its spreading of disinformation and the violation of the Twitter guidelines (Eustachewhich, 2020). This prevented the account from playing a larger role in the spread of disinformation.

4.5.3 Platform Shutdown

This involves shutting down an entire (social) online platform. This often is a legal request, a notice, and takedown request, from Police Authorities. This is much more difficult to do since the platform or country of origin might not collaborate. However, it can be achieved by collaborating with law enforcement in the country of origin and when possible, servers can be seized to gather evidence on the illegal activities so that the content owners and other related culprits can be prosecuted. An example of platform shutdown is the shutdown of Parler. Parler is an American social networking service which is highly associated with the conservative movement in the U.S. It marketed itself as a free speech focused and unbiased alternative to mainstream social media networks (Parler, 2025). The platform was suspended by Google, Apple and Amazon because it did not comply with their user guidelines of hate speech, as the platform hosted a multitude of vile disinformation and threats of violence against minorities (Thorbecke, 2021). The platform was eventually allowed back online, but lost a lot of followers and did not fully recover.

4.5.4 Platform Suspension

Platform suspensions entail temporarily blocking or disabling an entire (social) online platform. This often starts as a legal request from Police Authorities to shut down a website or platform, leading to a legal entity such as judge or the online platforms or service providers themselves deciding to suspend the platform until further notice. Mostly this is part of a criminal investigation and can be done in collaboration between countries. Platform suspension can also extend to intervening with

legal action against (for-profit) "factories" creating misinformation (DISARM). Often, they are aimed at creating money out of engagement in certain platforms or receive funding from others. Political influence or the power of states, including law enforcement efforts, can stop the entities behind critical outlets online, through content take downs or even removal of platforms. This goes beyond removing a platform or content but is aimed at the organisation behind it. An example of this is the United States blocking the social media platform TikTok, because of the platform not complying with U.S. Law. The platform was inaccessible for a short period of time, as after 12 hours the ban was reversed by a promise of Donald Trump (Dedezade, 2025).

4.5.5 Deleting Bots

This involves a coordinated effort to delete fake accounts and bots from social media, in order to limit falsified content and instigations. These requests can be done by Police Authorities and even end users (for example in case of impersonation), however a lot of social media platforms try to detect and remove these accounts themselves. This can also include the reduction or even blocking of access to the means of spreading disinformation, such as (requesting to) remove it from search engines or search options in social media platforms (DISARM). Police Authorities can both collaborate with service providers and enforce them to deal with malevolent content. Most community standards or policies of online platforms have rules that forbid impersonation or the use of accounts that are not controlled by real people.

4.6 Category 6: Police Mandated Actions

There are actions that can only be undertaken by Police Authorities, as they are the only ones with mandate to execute certain actions. Examples include arresting individuals or using legal procedures to track (online) information. The interventions below are ones that can be specifically applied by Police Authorities, due to police policies or police mandates. This could differ per country, but overall, these interventions are based on a global understanding of Police Authorities.

4.6.1 Arresting Miscreants

Arresting miscreants entails arresting those who are in obvious violation of a set behavioural code of online conduct. It differs per country what type of (online) behaviours are considered illegal by criminal law but related to disinformation and the different impact models used in VIGILANT. Examples include various forms of cybercrime (such as online fraud), specific forms of instigation of violence and organising (illegal) mass gatherings and various forms of hate speech. In Pakistan, a man was arrested on the basis of his fake accounts which he used to spread disinformation about the Southport attack and subsequent riots in the U.K. (Davies, 2024). It is, however, important to note that arresting people on the basis of spreading mis- and or disinformation, can be very subjective. In non-democratic countries, the definition of fake news and disinformation has been expanded in such a way that also encompasses any criticism of the regime (Mahapatra, 2024). This intervention needs to be monitored and carried out with the utmost caution.

4.6.2 Aware of Police Presence

This intervention refers to making individuals conscious that the Police Authorities are aware of their (online) presence by giving certain online and offline signals and communication (campaigns). Online signals could include replying under posts or direct messaging. The idea is that if surveillance or presence is communicated in risky environments individuals could appeal to this form of (policing) authority, pay more attention to their deviant behaviours and be more mindful of (not expressing) extreme behaviours that could cause harm or danger. Research has found that online police presence can mimic encounters between the general public and the police, thus influencing the behaviour of those interacting with the authorities, but that there is difficulty in guaranteeing the authenticity of a police account (Henry, 2024). However, the role of Police Authorities online is precarious. In 2018, Bellingcat looked into the Dutch Police activities online, and found that they were not careful in handing privacy of the people they interacted with online (Mulder, 2018). This can lead to a significant downfall of trust in Police Authorities and violate constitutional rights.

4.6.3 Cease and Desist Conversation Individual (Online)

A (warning) conversation as a preventive measure in order to try and reverse their path towards more extremist and dangerous or criminal behaviour. This could concern both first offenders, especially youngsters, but also repeated offenders where it is known to have effects on their possible future behaviours. This is used in cases of misdemeanours and small criminal offenses. These conversations could be either by voice communications or direct messages. It is known to have an impact to deanonymize a person in a larger online group conversation by reaching out to them and having personal contact with some form of authority, often resulting in voluntary removal of content and expressing excuses as most people are unaware of the impact of their online behaviour (especially young people). Public conversation (when private conversation is not possible) could spiral into perceived public shaming of individuals and have backlash from others joining in and is not considered effective nor ethical.

4.6.4 Cease and Desist Conversation Group (Online)

A (warning) conversation as a preventive measure to reverse the path of (potential) culprits, not targeted at stopping or preventing further harmful behaviour of an individual but to have effect on a whole group. This is used in cases of misdemeanours and small criminal offenses committed by multiple people in the group. There are several ways this can be done, such as replying to individual messages everyone in the group can see or joining the conversation altogether explaining the impact of their behaviour, giving a warning with criminal charges that might come into play if they continue and making them aware of police presence or making others aware they can report these matters (such as hate speech) to police if it impacts them.

4.6.5 Cease and Desist Conversation Individual (Physical)

A (warning) conversation as a preventive measure in order to reverse the path of (potential) culprits. This is used in cases of misdemeanours and small criminal offenses, for example in cases of stalking. This is often done through a personal one to one

conversation, sometimes at the house, school or working location of a person. In order to be able to do this, identification of the person needs to be possible, often done with a user account profile that contains personal information or if a person is already known to police as a repeated offender. If it concerns a minor, it is desirable that a parent or other legal guardian is present at such a conversation. This intervention is based on internal police documents which are restricted to public access.

4.6.6 Cease and Desist Conversation Group (Physical)

A (warning) conversation as a preventive measure to reverse the path of (potential) culprits, not targeted at stopping or preventing further harmful behaviour of an individual but to have effect on a whole group. This is used in cases of misdemeanours and small criminal offenses committed by multiple people as part of a specific group, such as people that are part of a local community group, a soccer team, school class or any other form of social cohesive group that can also be addressed in the physical world, with contexts such as sports, school, work or streets. Examples could include a preventive talk of a police officer at a streetcorner where youngsters hang out or at a school in front of a whole classroom if more are involved, or in a more private setting such as a conversation with a specific group or people together with a dean and parents. This intervention is based on internal police documents which are restricted to public access.

5 Characteristics of Interventions

All interventions were categorized based on their characteristics. Several distinctions were made to determine the relevance of each intervention to specific situations or PAs. The necessary characteristics for usefulness are described below. The selection and description of these characteristics were primarily based on expert opinions of the research team, on literature, on interviews with PAs, and considering the scope of the VIGILANT project.

5.1 Types of Intervention

5.1.1 Physical or Online

An intervention or counter measure for disinformation can be done online, in the real physical world, or both. In practice most social media platforms can or will only intervene online and most policing organisations have experience intervening in the offline or physical world. Sometimes more effort and additional considerations are needed to combine the online and offline worlds for an intervention to be (more) effective or to measure the effects. It can also be more beneficial to do both simultaneously, depending on the context and (online or offline) impacts of the disinformation, for example disinformation that is spread in both the online and offline world.

5.1.2 Preventive and/or Repressive

Preventive measures are aimed to prevent, minimise, or mitigate negative impacts such as harm or damage. These measures are taken to prevent disinformation from starting at all, or to prevent disinformation from having an effect. A prevention of disinformation starting involves automatically blocking certain content (which raises censorship issues). One example of an effective preventions of disinformation is inoculating people so they are resilient to disinformation.

Repressive measures are methods employed by Police Authorities to suppress disinformation from spreading further. The disinformation or other type of harmful content is already ‘out there’ and could potentially have violent repercussions. This includes removing content in violation of the code of conduct or tracing instigators for a cease-and-desist conversation.

5.1.3 Escalation Level

The escalation levels are an indication of the severity of societal impact, as some interventions are more suited in early stages of (smaller) societal impacts and other interventions are more suited for more severe or broader impacts. In the case of mass gathering these levels are described in Section 4.3 of D5.1: Impact analysis tool (Maas et al., 2024).

5.1.4 C5 Model

The C5 model explains the different aspects (Context, Causes, Content, Cycle of Amplification and Consequences) that could all in some way or form play a role in the spread of disinformation. Interventions could apply to more specific aspects, multiple

or all of the C's as explained in Section 4.2 of D5.1: Impact analysis tool (Maas et al., 2024). A more detailed description of the C5 model can be found in Section 4 of D2.4: Causes, contents and consequences model (Kruijver et al, 2023).

5.2 Timeframe

5.2.1 Timeframe Development

Interventions take time to develop. Either the whole method needs to be tailored to a situation, or the contents of that method (such as a counter narrative) should be tailored to be fit for its purpose which, depending on the context, can take hours, days, weeks etc. The timeframe in the Disinformation Intervention Framework is defined in terms of hours, days, weeks or even months.

5.2.2 Timeframe Execution

Operationalising or executing an intervention also takes time. Some interventions, such as information campaigns, can take months, whilst other interventions can be executed and completed on the same day, such as arresting a miscreant. For other interventions, the execution itself could be done in a swift manner, but often takes weeks or even months before the process is completed. Only in crisis or emergency situations are these such interventions sometimes sped up because of the legal authority of PA,s or because platforms have special procedures for cases that involve, for example, terrorism. Removing content or legal requests from police for IP addresses to help in identifying an online user can take weeks or sometimes just hours depending on the context.

5.2.3 Timeframe Effectiveness

This is the estimated duration for an effect of a given intervention. Some interventions have short effect durations, while others may have permanent effects. For example, a counter messaging campaign may only have an effect on people's perception of the disinformation for a number of weeks, while shutting down an account has a permanent effect on the disinformation output of that account. For some interventions, such as those relating to resilience, the duration of the effects is hard to estimate since they are dependent on a multitude of mitigating factors.

5.3 Execution

5.3.1 Intervenor (Owner/Starter)

Who is the main intervenor in this intervention? In the execution, it helps speed up the process if police can perform the whole intervention themselves. For example, creating an online counter narrative or posting a reaction might be something they can do and decide upon by themselves. Sometimes police are also owner of the whole intervention, such as handing out flyers on the streets. In the case of online policing, they may have their own website but on other platforms, PA users are solely end-users of the platform, like any other end-user and are very much dependent on the decisions and restrictions of that platform.

In some cases, such as in criminal cases where, as arrest seems imminent, police depend on decisions by others because in most countries police may not be able to act without a decision from public prosecution.

5.3.2 Mandatory Intervention Partner

Who is needed (essential) for the intervention implementation? In some cases, police must rely upon partners in the execution of an intervention. For example, if the reason for an intervention is based on suspected illegal actions, a collaboration with a prosecutor may be necessary. Also, in some cases of deplatforming, a PA is not able to delete content themselves but may need to put in a request with a service provider for them to take down content.

5.3.3 Possible Intervention Collaborator

In other cases, third parties are a 'nice-to-have' and could help improve the impacts of an intervention. An example of this is a social media campaign aimed at a certain target group. Social media platforms, and other online service providers, NGO's or education partners, could help provide the means to boost this message and make sure it targets the right audience. On social media platforms, an example is the current practice of online advertisement where police could have specific requests on these advertisements, such as asking platforms to prevent certain ads from being placed in order to prevent social unrest. An example of this is an advertisement for a chainsaw next to the posts of a murder case where people are asked to call police if they know more.

5.3.4 Target Audience of Intervention

Defining and understanding the target audience for your intervention is vital to its success. Not all interventions are suited for all, therefore this report tries to describe if they are suited for a general audience, or only suitable for perpetrators or other specific audiences.

5.3.5 Targeted Outcome of Intervention

The targeted outcome is the goal of the intervention. Examples of target outcomes of interventions include raising awareness, removing root causes or containing the spread. It could also include the establishment of a normative shifts (in a group, platform, region or even broader society), engaging others in the intervention, putting up barriers or deterring individuals or preventing secondary effects (e.g. social, economic, political). There may also be ancillary effects (e.g. polarisation, curtailing freedom of speech, unfairness).

5.3.6 Scalability of Intervention

This concerns the extent to which an intervention by Police Authorities can be scaled up. Some interventions will start small with the possibility of escalating to a higher scale, for instance by applying to individuals first and then being rolled out to target a wider group. However, this scalability depends on the capacities and resources of the police. For example, online or digital

interventions are often easier to scale than physical interventions because certain actions can be automated, particularly if it is spread on just one or a few big platforms.

5.4 Considerations

These considerations of the interventions are possible limitations to consider before implementing an intervention. Particularly as ethical considerations are concerned (see Section 2.4), it is important to take note of the effects that interventions may have. Interventions may backfire, but also interventions used jointly might interfere with each other.

5.4.1 Judicial Framework or Policy

What are the relevant legal frameworks or policies that might apply to this intervention? This concerns any array of legal or policy constraints, including social norms or (perceived) proportionality (legitimacy) and ethical considerations, including cultural differences, which may impact the ability to implement interventions. For example, there may be certain criminal laws, constitutional rights, and EU regulations that limit what authorities can do online. Furthermore, Police Authorities themselves may be bound by their own policy or regulatory frameworks regarding online activity.

5.4.2 Risks of Intervention

What are the limitations of the intervention in terms of risks for the police and other stakeholders? This concerns, for example, potential harmful side-effects or criticism that impacts police or government reputation or a countermeasure that could entail some backlash effects. For instance, there is a risk of waterbed effects (e.g. the intervention only acts to move the threat to another platform or domain), censorship, violent repercussions, lowering trust in police, polarisation, reputational damage and infringing norms and values. This list of examples is not exhaustive and there may be many more risks associated with interventions online.

6 Discussion

This report has sought to outline the Intervention Support Tool and its constituent parts, most importantly the Disinformation Intervention Framework. Although the underlying framework provides a comprehensive overview of disinformation interventions and their characteristics, it is not exhaustive. Furthermore, the framework provides a comprehensive set of intervention supports based on their potential effectiveness in countering disinformation. When attempting to counter disinformation online, however, there are inherently a series of constraints and potential risks that need to be considered when deciding what intervention best fits a given situation, especially since this framework applies to PAs across the EU. The interventions outlined in this report all come with certain specific caveats or conditions which may influence their design, implementation or effectiveness. For expediency, this section outlines the main considerations that Police Authorities must take into account when designing and implementing interventions to counter disinformation. These are grouped into four sections: considerations about effectiveness, legal considerations; ethical and social considerations; and procedural considerations.

6.1.1 Effectiveness Considerations

Even if all other considerations are satisfied, there remains a question of effectiveness. The fundamental question here is whether the police are the most effective authority to intervene or are there other more effective implementers? Social media platforms retain the most significant digital intervention experience, while the expertise of Police Authorities still lies primarily in physical, real-world interventions. The power of other online actors, such as influencers, should also not be underestimated, particularly if an intervention is aimed at targeting young people. These variations in expertise and ability to implement interventions should be actively considered as it may be the case that influencers or social media platforms are far more effective implementers of an intervention than the police. A solution to bridge this gap and ensure effectiveness is to improve collaboration between law enforcement authorities and social media platforms.

Another key consideration in relation to effectiveness is to not be overly reliant on the results of detection and monitoring, which can be misleading. As discussed above, data-driven and automated analyses provide law enforcement with fast, large-scale analyses but there is always the concern of internal issues over biases, false assumptions and profiling of audiences in the tooling. Therefore, to ensure that any online police activity is effective and accurate, it is essential that human officers retain a central role in the process, whether in detecting, monitoring or mitigating disinformation.

6.1.2 Legal Considerations

The most obvious constraint that Police Authorities face when countering disinformation are the legal limits placed on their activity online. Across Europe, mass unfettered online surveillance by police conflicts with basic constitutional and human rights, as the EU protects individuals' personal data used by law enforcement in their prevention, investigation, detection and prosecution of potentially criminal offences. Any personal data that is collected must comply with certain principles founded

in EU law (such as the GDPR) or national legislation (on media or other relevant issues), and specific legislation for PA's, including that the data is processed lawfully, collected for specific, explicit and legitimate purposes and it is not excessive.

Even if the collection and use of personal data online satisfies these conditions, any response or intervention is bound by the policing principles of proportionality and necessity. There are also important considerations over the targets of any police activity online. When designing interventions for PAs online, it is important to consider whether the activity is targeting minors, those aged 18 and under, or whether it is disproportionately targeting other vulnerable individuals, marginalised groups or minorities. Furthermore, there is a risk that any police intervention may impinge on the constitutionally enshrined right to freedom of expression and association. Taken together, these legal constraints on police activity in the detection and intervention of disinformation are of utmost importance when designing an appropriate course of action.

6.1.3 Ethical Considerations

Countering disinformation online can produce other unwanted and potentially harmful effects which are ethical in nature, rather than legal. A primary concern here is transparency and accountability. Although this is important when monitoring and detecting disinformation, it is a particularly important consideration when Police Authorities design and implement countermeasures and interventions against disinformation. Furthermore, if collaboration with social media platforms is required for an intervention, which is often the case, it is important to bear in mind that there may be a conflict over who claims credit, the police or the platform. Whether the police are attributed to the intervention, and the extent to which this is made public, is a fundamental ethical consideration because of the public role of law enforcement and the importance of maintaining trust in police services. Any inclination among the public that police activity online is underhand, unlawful or not fully accountable could lead to ethical dilemmas or reputational damage for the police, which could lead to an undermined authority. Overall, the perception of the intervention is often just as important as the intervention itself. As such, it is advisable that a 'rule book' on transparent and accountable police activity is publicised and adhered to.

Another ethical consideration is the profiling of audiences, both in terms of disinformation detection and intervention. There may be legal limitations to this (see Section 6.1.2) but there are also ethical concerns, even if the law is abided by. The police should not seek to profile individuals based on group characteristics or demographics alone, particularly if that profiling identifies them as part of a minority, vulnerable or marginalised group (social, ethnic, religious, political etc.). For instance, targeting adverts to specific groups online based on demographics will likely constitute as unethical profiling. This can be difficult to avoid given that any social media intervention carried out by the police will probably require the large-scale use of data-driven technologies and automated analyses. However, it is important to ensure that any biases are alleviated. In summary, Police Authorities should be wary of algorithms in online analyses and retain a significant role for human effort and cognition.

A more social or political consideration, rather than purely an ethical one, is that of the second-order effects of online interventions against disinformation. Owing to some of the dynamics listed above, a heavy-handed approach by the police in

detecting and countering disinformation online may lead to damaging, harmful or undesirable consequences. An overly stringent and vigilant online police presence may achieve the desired effect in the short-term, but it could elicit a backlash against police surveillance, or it may simply drive users to use other, less regulated platforms. Both have far-reaching consequences. For example, although not an example of a police intervention, the banning of Donald Trump on Twitter received vociferous criticism from his supporters who followed him to his own social media platform, Truth Social. Rather than reducing and preventing the spread of disinformation, there is evidence to suggest that this intervention enabled greater polarisation within U.S. political discourse (Alizadeh et al., 2022). More detail on ethical considerations for the interventions as part of the VIGILANT platform as a whole can be found in D2.1 (Ethics Framework), D2.2. (Ethics guidelines for Police Authorities) and D2.3 (Ethical oversight report).

6.1.4 Procedural Considerations

Police procedures are unique in how they can potentially put constraints on the detection and mitigation of online disinformation. The first, and most important of these, is whether the online content in question constitutes a police matter. In essence: is the disinformation criminal or sufficiently related to criminal activity? Although somewhat obvious, this distinction is not always clearcut and requires nuanced consideration. Another consideration is whether police intervention against disinformation interferes with other police processes. For example, shutting down certain social media accounts may impede intelligence collection. Similarly, police interventions may interfere with or conflict with social media platforms and other partners. Many of the interventions listed in this report require collaboration between the police and other partners, such as social media partners. Therefore, the rules, policies and capabilities of platforms must be taken into account when developing interventions. The use of automated content moderation varies across platforms¬—for example, TikTok uses algorithms to automatically remove content —and so this must also be considered when designing interventions. It is important to note here that police interventions against disinformation are not simply a ‘push of a button’ but require considerate planning and considerable collaboration, whether with platforms or other authorities.

6.1.5 Collaboration Considerations

An important element of the development of the Intervention Support Tool was collaboration and integration. Through collaborating with several PAs, the authors obtained insights about their operations and the tools and knowledge on disinformation interventions that are currently available to them. Through workshops and interviews (see Section 3), PAs indicated that sharing all the available information and specific insights from online groups is not allowed, since these are labelled as classified police information. However, sharing the knowledge about how effects on an individual level can escalate to effects on a group level is helpful in increasing common understanding between police and (local, regional or national) public partners. Such knowledge sharing can also help to create more awareness within these organisations and help raise awareness among the public at large. The PAs acknowledged that many cases of disinformation are not restricted to national borders or boundaries of their policing tasks. Therefore, intra-regional or international collaboration with other PAs and Europol/Interpol is an important consideration when developing intervention options.

Other potential collaborative public partners might, for example, be municipalities, national governments, social workers, schools or NGOs that can all play a role in countering disinformation. These organisations typically tend to use preventive instead of repressive measures, which are mostly considered a last resort and part of the police tasks. Coordination between these partners on their interventions by sharing their insights using the knowledge and insights from VIGILANT could assist police organisations to be more effective and efficient overall.

7 Conclusions

This report describes the Disinformation Intervention Framework that was developed to support the decision-making of Police Authorities when carrying out offline and online interventions related to disinformation. It lists and describes interventions divided in six categories to provide a comprehensive overview for decision makers within PAs to be more effective in countering disinformation. All these interventions are built into the VIGILANT platform as part of the Impact Analysis Tool (see D5.1: Impact analysis tool, Maas et al., 2024).

This report also describes the science and evidence base behind many interventions, although it must be noted that little is known on the actual effectiveness of interventions, both within police context and in scientific literature. Therefore, several examples of real-world practices are given and, if relevant, legal frameworks are highlighted. In order to make proportional and effective decisions, the theory-based C5 model has been integrated into the Disinformation Intervention Framework and the Intervention Support Tool (D5.2). Furthermore, ethical and legal considerations (as addressed in WP2: Ethics, Disinformation and Requirements) are integrated into the work.

7.1 Limitations & Challenges

One of the main limitations of this report, and the task of T5.2: Intervention support tool, is that the framework of disinformation interventions, and their characteristics, is not exhaustive. Indeed, this report serves to provide an overview of the most prominent and prescient interventions today and the authors acknowledge that other intervention options that are not mentioned in the above may exist and prove fruitful. Similarly, the list of considerations (Section 5.4: Considerations) is also non-exhaustive and serves to provide an initial outline of the main legal and policy constraints as well as potential risks associated with any given intervention.

This relates to two further limitations of this report. First, many of the interventions and their characteristics are dependent on context. In other words, the jurisdiction in which each Police Authority operates might alter the applicability of certain interventions and their potential risks. For example, the effectiveness of an intervention (Section 5.2.3: Timeframe effectiveness) may vary greatly depending on the context. Second, the longevity of the framework described in this report is limited. Due to the ever-changing nature of online activity, particularly the nature and dissemination of disinformation, it is possible that the framework described in this report may become outdated rather quickly. Therefore, a certain level of intuition and flexibility is required by Police Authorities in interpreting and applying this framework to their own work and jurisdiction in the future.

Although the work of T5.2: Intervention support tool is described in whole in this document and can be demonstrated as a working component as part of the VIGILANT system, there are still in practice some gaps and interrelations that have to be filled to integrate the intervention support tool successfully into the entire VIGILANT system. As technology progresses new improvements and additions will be made after this phase adding new analytical components, where integration of all

components will need attention. In the VIGILANT solution the Intervention Support tool (D5.2) was developed as a part of the Impact Analysis Tool (described in D5.1: Impact analysis tool). Both the impact analysis component and the intervention support component already make use of most analytical components, so as new components are added integrating them needs implementation efforts using the flexible general architecture described in D6.1: System architecture.

Alongside the next phase of validation and exploitation, before introducing any new system or way of working the training from WP7: DC&E (Dissemination, Communication and Exploitation) and Training, will have to be successfully implemented. And before introducing any new system there are procedures within all PAs to carry out ethical and legal assessments. Examples of these assessments include law enforcement legislation in different EU countries, guidelines and relevant EU legal frameworks such as the GDPR, the AI Act and Digital Services Act for service providers and social media platforms that PAs collaborate with.

VIGILANT will contribute to the scientific understanding of the underlying social drivers motivating crimes or misdemeanours related to disinformation and other related issues such as hate speech, radicalisation, extremism, and violent separatist movements. It will further translate this knowledge into guidelines for tools, eligible for PA use. Specifically, the Impact Analysis Tool (D5.1) and the Intervention Support Tool (D5.2) will enable PAs to target responses to prevent (or stop) disinformation taking hold in a community. Long-term, this will lead to deterrence of perpetrators of these crimes, provided that the investigation capabilities of PAs will increase as well as the cross-border cooperation). Secondly, having PAs from four different countries and the Early Adopter Community participating in the project and contributing to the development of the platform and the training materials will foster cross-border cooperation and sharing of best practices among the involved PAs and the member states in general. This work will contribute to the VIGILANT training and exploitation phase (T7: DC&E (Dissemination, Communication and Education) and Training), where the link is made to operational or tactical police practices. Recommendations from the D5.1 and D5.2 reports and T2: Ethics, Disinformation and Requirements can be taken a step further in the training and exploitation phase.

Below, some further observations regarding the use of these interventions in practice and more research opportunities are discussed. The deliberations, which include conclusions as well as limitations of this work, can be categorised along two lines: the foundation in state of the art scientific and ethical/legal developments on the one hand and its contribution and relevance as an innovative intervention support tool to police practice with regard to disinformation on the other.

7.2 Recommendations for Future Development, Implementation and Uptake of the VIGILANT Disinformation Intervention Framework and Intervention Support Tool

Based on the work done in the VIGILANT project and specifically within WP5: Social Drivers and Behavioural Dynamics, we found several recommendations for research on and prevention of disinformation. These recommendations vary from using the ever-improving technology to enhance detection and to further develop human interaction with technology, to increase

collaboration between Police Authorities and, finally, to implement the work into a broader context. The fight against disinformation is a multidisciplinary one, and therefore, many factors should be considered when continuing this work.

7.2.1 Technology and Human Interaction

The Impact Analysis Tool that is described in D5.1 highlights how responses to disinformation cannot and should not be based on assumptions or intuition. There is no shortcut or automatic quick fix for disinformation cases or campaigns. There are a lot of considerations such as assessing the context, the underlying dynamics and potential second order effects that might be undesirable. So before deciding on any intervention a lot of help is offered on procedural, legal, ethical and societal considerations that might differ per country and organisation. One way of offering more help is to include the recommendation of best practices, another is to not rely on the technology alone to decide in isolation but rather to always consult peers within or outside of your organisation before an action is taken.

Lastly, to prevent some of these pitfalls, the training of PAs, if possible, together with partners, in countering disinformation is crucial. Experimenting and discussing both analyses and possible actions can help foster a reliable and accountable Police Authority in countering disinformation that helps build a legitimate position for PAs as trustworthy and important partners with a unique and crucial role in the countering of disinformation.

In the exploitation phase, after the project, which is described in D7.4: DC&E (Dissemination, Communication and Education) and Training, there should be enough emphasis on the 'human in the loop' at several stages from start to finish before real world actions are taken.

7.2.2 Collaboration Between PAs and with Other Partners

The online domain, including a number of social media platforms and information environments, is constantly shifting and changing at a fast pace, faster than science or legislation can accommodate for. Therefore, police practice will always have to deal with new and unexpected interventions or actions, where existing legal or ethical framework may not be sufficient. Criminals, and also other internet users, can be very innovative and as new technologies arise, they can find new opportunities in countering disinformation. The categories and list of interventions that can be found in this report is a good starting point, and are supported by science, by examples, and legislation and ethical considerations. It is recommended that knowledge is shared on these matters within and across Law Enforcement Agencies, with other government agencies, as well as NGOs and private partners, in order to continue to evolve by sharing best practices. This can be done in EU bodies such as Cepol, Europol or other international collaborations related to disinformation. Sharing knowledge models can be a means to support this and in some international cases expanding them towards international models for social networks that are interacting internationally can provide insights that so far have been limited to some individual cases.

Not only are public partnerships important, but private partners can play a role in both the online and offline domain, where specific target groups can be reached with preventive interventions. Internet service providers, (local, regional or national), media partners (online or offline), service providers and any other private businesses could either be impacted themselves by disinformation should play an important role in countering disinformation. This presents opportunities for PAs to have a broader reach with the information they gather and will increase the opportunities for fruitful interventions.

Besides international collaboration between PAs and other public partners, improving the collaboration with Big Tech companies that operate internationally is important too. These include the largest international social media platforms, and also gaming platforms and international media outlets. Again, sharing knowledge and insights using these impact models and their structure, could help a great deal to be more effective and efficient in countering disinformation. The models could be improved by adding 'pieces of the puzzle' as only all partners contributing together can create a comprehensive overview of certain online or offline phenomena. The private (social) platforms often see more online content than public partners can, while public partners are often more engaged in the physical world with end-users than these Big Tech partners are. By working together, they can coordinate their efforts and provide a more integral approach that could effectively make a difference.

It is not always clear who is responsible for which actions and although Police Authorities have a role in the deterrence and prevention of crimes and harm related to disinformation, they also very much depend on their public and private partners to make a difference. Citizens can also play a role in reporting or even countering disinformation, as the problem grows in both online and offline media. Therefore, a recommendation would be to create a clear owner of the problem and give this owner an appropriate mandate to handle the issue of possible harmful information.

7.2.3 Embeddedness in Broader Policing Context

It is also important to note that traditional policing, such as starting a criminal investigation or arresting someone, is more of a final frontier in countering disinformation. On the other hand, Police Authorities could help more in the prevention of disinformation or countering it in early stages, since they have a unique information position and sometimes are allowed to detect more in the online and offline information environment. In this manner police could share information where possible (not necessarily personal information, but information on new phenomena or modus operandi) with public or private partners and also receive information from these partners so all can do a better job at countering disinformation and prevent harm.

7.3 Future Work

During the project both within the consortium and talking with stakeholders outside of the consortium, several suggestions have been expressed as to how technology could further assist police analysts in making sense of disinformation, particularly as the topic is complex, and involves the emergence of new technologies and online contexts.

One such possibility that is currently discussed is the use of the Intervention Support Tool in training simulations to teach the concepts of disinformation and how to fight it. This would be an extension to the training methods developed in WP7: DC&E (Dissemination, Communication and Education) and Training.

Several considerations for an exploitation phase have also been discussed in Section 7.1. These considerations, such as improved collaboration with other partners or EU projects, the inclusion of more PAs and a maintained focus on ethics, are important aspects in all future work to ensure that the eventual tool can and will be used by PAs across Europe.

8 References

Alizadeh, M., Gilardi, F., Hoes, E., Klüser, K. J., Kubli, M., & Marchal, N. (2022). Content moderation as a political issue: The Twitter discourse around Trump's ban. *Journal of Quantitative Description: Digital Media*, *2*.

Allyn, B. (2020). Twitter Removes Thousands Of QAnon Accounts, Promises Sweeping Ban On The Conspiracy. *NPR,* July 21st 2020. https://www.npr.org/2020/07/21/894014810/twitter-removes-thousands-of-qanon-accounts-promises-sweeping-ban-on-the-conspir

Baraz, A., & Montasari, R. (2023). Law enforcement and the policing of cyberspace. In *Digital Transformation in Policing: The Promise, Perils and Solutions* (pp. 59-83). Cham: Springer International Publishing.

Bateman, J., & Jackson, D. (2024). Countering disinformation effectively: an evidence-based policy guide. *Carnegie Endowment for International Peace.* https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en

Bromell, D. (2022). Deplatforming and democratic legitimacy. In *Regulating free speech in a digital age: hate, harm and the limits of censorship* (pp. 81-109). Cham: Springer International Publishing.

Bell, Karissa. (2023). Instagram adds false information labels to posts. Mashable. Retrieved March 14, 2025, from https://mashable.com/article/instagram-false-information-labels

Business & Human Rights Resource Centre. (2023). "Apple and Google disable maps features in Israel and Gaza*. Business & Human Rights Resource Centre." Retrieved March 14, 2025, from https://www.business-humanrights.org/en/latest-news/apple-google-disable-maps-features-in-israel-gaza/

Bryanov, K., Vasina, D., Pankova, Y., & Pakholkov, V. (2021, June). The other side of Deplatforming: right-wing telegram in the wake of trump's Twitter Ouster. In *International Conference on Digital Transformation and Global Society* (pp. 417-428). Cham: Springer International Publishing.

Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, *18*(1), e12890.
Cookson, D., Jolley, D., Dempsey, R. C., & Povey, R. (2021). A social norms approach intervention to address misperceptions of anti-vaccine conspiracy beliefs amongst UK parents. *PLoS One*, *16*(11), e0258985.

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, *28*(11), 1531-1546.

Community Notes. (n.d.). Introduction. Community Notes Guide. Retrieved March 14, 2025, from https://communitynotes.x.com/guide/en/about/introduction

Davey, J., Tuck, H., & Amarasingam, A. (2019). An imprecise science: Assessing interventions for the prevention, disengagement and de-radicalisation of left and right-wing extremists. *Institute for Strategic Dialogue*. https://www.isdglobal.org/isd-publications/an-imprecise-science-assessing-interventions-for-the-prevention-disengagement-and-de-radicalisation-of-left-and-right-wing-extremists/

Davies, Caroline. (2024). Pakistan arrests Man over Southport Attack Disinformation. 21 August 2024, BBC News. https://www.bbc.com/news/articles/c05je6yz0q1o

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29.

Ecker, U. K., Sanderson, J. A., McIlhiney, P., Rowsell, J. J., Quekett, H. L., Brown, G. D., & Lewandowsky, S. (2023). Combining refutations and social norms increases belief change. Quarterly Journal of Experimental Psychology, 76(6), 1275-1297. https://doi.org/10.1177/17470218221111750

Eustachewhich, L. 2020. Twitter locks Trump campaign's account over video related to The Post's Hunter Biden bombshell. *New York Post,* October 15 2020. https://nypost.com/2020/10/15/twitter-blocks-trump-campaign-from-tweeting-posts-hunter-biden-story/

Evans, R. (2023, February 20) Police spy unit cause 'outrage and pain' as it infiltrated leftwing groups. *The Guardian.* https://www.theguardian.com/uk-news/2023/feb/20/police-spy-unit-caused-outrage-pain-infiltrated-leftwing-groups

Fortin, F., Delle Donne, J., & Knop, J. (2021). The use of social media in intelligence and its impact on police work. *Policing in an Age of Reform: An Agenda for Research and Practice*, 213-231.

Frenkel, S. (2020). Facebook Removes 790 QAnon Groups to Fight Conspiracy Theory. *New York Times*, August 19th 2020. https://www.nytimes.com/2020/08/19/technology/facebook-qanon-groups-takedown.html

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ data science*, *11*(1), 3.

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. Journal of Management Information Systems, 38(1), 196-221.

Grimmelman, J. (2015). The Virtues of Moderation. *Yale Journal of Law & Technology,* 42.

González-Bailón, S., & Lelkes, Y. (2023). Do social media undermine social cohesion? A critical review. *Social Issues and Policy Review*, *17*(1), 155-180.

Google Services. 2023. Search Quality Rater Guidelines: An overview. November 2023. https://services.google.com/fh/files/misc/hsw-sqrg.pdf

Hollywood, J. S., Vermeer, M. J., Woods, D., Goodison, S. E., & Jackson, B. A. (2018). Using social media and social network analysis in law enforcement. *RAND Corporation, Santa Monica, CA, USA*.

Hunt, K., Agarwal, P., & Zhuang, J. (2022). Monitoring misinformation on Twitter during crisis events: a machine learning approach. *Risk analysis*, *42*(8), 1728-1748.

In Cyber News (2024, June 11) Canada: RCMP plans to infiltrate online "extremist groups". *In Cyber News Editorial Team, Digital Transformation.* Available at: https://incyber.org/en/article/canada-rcmp-plans-to-infiltrate-online-extremist-groups/

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K., Lewandowsky, S., Hertwig, R., ... & Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044-1052.

Kruijver, K., Finlayson, N. B., Cadet, B. & van der Meer, S. (2023) Causes, contents and consequences model. *VIGILANT, D2.4,* EU Horizon. Available at: https://www.vigilantproject.eu/outputs

Kruisbergen, E. W. (2021). When Other Methods Fail…: Infiltrating Organized Crime Groups in the Netherlands. *Contemporary Organized Crime: Developments, Challenges and Responses*, 249-274.

Lai, L. S., & To, W. M. (2015). Content analysis of social media: A grounded theory approach. *Journal of electronic commerce research*, *16*(2), 138.

Laor, T. (2024). Breaking the silence: the role of social media in fostering community and challenging the spiral of silence. *Online Information Review*, *48*(4), 710-724.

Maas, Y., de Vries, A., de Jong, A., Finlayson, N. B. & Poot, E. (2024) Impact analysis tool. *VIGILANT, D5.1.* EU Horizon. [soon to be published] Available at: https://www.vigilantproject.eu/outputs

Mahapatra, S., Sombatpoonsiri, J., & Ufen, A. (2024). Repression by legal means: Governments' anti-fake news lawfare. GIGA Focus Global, 1, 11. https://doi.org/10.57671/gfgl-24012

McBride, M. K., Faber, P. G., Haney, K., Kannapel, P. J. & Plapinger, S. (2024) Evidence-based Techniques for Countering Mis-/Dis-/Mal-Information: A Primer. *Center for Naval Analyses (CNA).* https://www.cna.org/reports/2024/03/a-primer-on-countering-mis-/dis-/mal-information

Meta. (2019). Combatting Misinformation on Instagram. December 16, 2019. https://about.fb.com/news/2019/12/combatting-misinformation-on-instagram/b

Nygren, T., & Ecker, U. K. (2024). Education as a countermeasure against disinformation. *Psychological Defence Research Institute.* Lund University*:* working paper, 3. https://www.psychologicaldefence.lu.se/article/education-countermeasure-against-disinformation

OSINT Industries Team (2024, October 31) Social Media Intelligence (SOCMINT) in Modern Investigations. *OSINT* Industries. Available at: https://www.osint.industries/post/social-media-intelligence-socmint-in-modern-investigations

Parler. (2025). Parler. https://www.parler.com/

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213-229.

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1-10.

Seering, J., & Kairam, S. R. (2023). Who moderates on Twitch and what do they do? Quantifying practices in community moderation on Twitch. *Proceedings of the ACM on Human-Computer Interaction*, *7*(GROUP), 1-18.

Selim, H. A., & Popovac, M. (2024). Social cohesion in an online era: opportunities and challenges on social media. *Handbook of Social Media in Education Consumer Behavior and Politics*, 279-298.

Stoycheff, E., Liu, J., Xu, K., & Wibowo, K. (2019). Privacy and the Panopticon: Online mass surveillance's deterrence and chilling effects. *New media & society*, *21*(3), 602-619.

Thorbecke, Catherine. (2021). Amazon reveals violent content, death threats that led to Parler's suspension. *ABC News*, January 14th, 2021. https://abcnews.go.com/Business/amazon-reveals-violent-content-death-threats-led-parlers/story?id=75221495v

Wang, K., & Fu, Z. (2023). Content Moderation in Social Media: The Characteristics, Degree, and Efficiency of User Engagement. Journal of Social Media Studies, 15(2), 123-145