VIGILANT

VITAL INTELLIGENCE TO INVESTIGATE ILLEGAL DISINFORMATION

Deliverable 5.1 – Impact analysis 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: D5.1: Impact analysis tool
Work Package: 5.1 – Impact analysis tool
Lead Beneficiary: TNO
Due Date: 31/10/2024
Revision Number: V1.6
Authors: Yuri Maas, Arnout de Vries, Arnoud de Jong, Neill Bo Finlayson, Elisabeth Poot
Dissemination Level: Public
Deliverable Type: Report

Overview: The purpose of this document is to explain and showcase the impact analysis tool as part of the VIGILANT platform. The analysis tool is based on impact models inspired by police escalation levels and the causes, contents and consequences model from deliverable D2.4 of the VIGILANT project. The document contains a developed concept for impact analysis, which was applied to a use case using example models that can be used in the greater VIGILANT framework.

Revision History

Version #	Implemented by	Revision Date	Description of changes
V0.1	Yuri Maas	14/06/2024	Document layout and styles
V1.0	Yuri Maas, Arnout de Vries, Arnoud de Jong, Neill Bo Finlayson	13/09/2024	Finalized first draft of the document chapter 2 to 9
V1.1	Arnout de Vries	2/10/2024	Executive Summary
V1.2	Elisabeth Poot, Yuri Maas	4/10/2024	Processing internal feedback; Added acronyms to table; Formatted Images
V1.3	Arnout de Vries, Yuri Maas	8/10/2024	Updated executive summary
V1.4	Elisabeth Poot, Yuri Maas	25/10/2024	Processing external feedback
V1.5	Arnoud de Jong	29/10/2024	Modified several figures for clarity
V1.6	Arnout de Vries, Yuri Maas	30/10/2024	Final tweaks for clarity and formatting across the document

Approval Procedure

Version #	Deliverable Name	Approved by	Approval Date
V1.0	D5.1	Yori Kamphuis, TNO Kimberley Kruijver, TNO	27/09/2024
V1.2	D5.1	Eva Power, TCD Ahmad Zareie, USFD	10/10/2024

Acronyms

The following table provides definitions for acronyms and terms relevant to this document.

Acronym	Definition
AI	Artificial Intelligence
D&FN	Disinformation and Fake News
DBKF	Database of Known Fakes
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 Organization for Standardization
IT	Information Technology
KInIT	Kempelen Institute of Intelligent Technologies
KPI	Key Performance Indicator
MCDA	Multi-Criteria Decision Analysis
MX	Month
NGO	Non-governmental organization
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

1. Executive Summary

The VIGILANT project focuses on developing an integrated platform to help police authorities (PAs) identify, analyse, and counter disinformation using advanced methods and tools, including the use of artificial intelligence (AI). The platform addresses disinformation across different formats (text, image, video) and languages, gathering information from major sources such as social media and (fake) news websites. It aims to enhance the ability of PAs to investigate or prevent unwanted societal impacts including criminal activities linked to disinformation.

The VIGILANT project collaborates with other EU projects like FERMI and DISARM, focusing on different approaches to tackle disinformation, allowing knowledge exchange to avoid pitfalls and improve methodologies. The project emphasizes ethical considerations, ensuring that its methods and (AI) tools respect privacy, avoid bias, and operate transparently whilst maintaining accountability through human interaction.

Work Package 5 (WP5) plays an important role in the project, with the objectives of understanding the social drivers and behavioural dynamics behind disinformation (WP5.1) and developing tools for selecting interventions for PAs and their partners (WP5.2). This includes the creation of an impact analysis tool that assists PAs in identifying, analysing, and responding to disinformation using dynamic system maps and indicator-based risk analysis. The tool applies conceptual models, which analyses 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.

The impact analysis tool is tested in real-world use cases, such as tracking hate speech from right-wing extremism in Catalonia and demonstrating its practical application. The tool helps analysts assess disinformation campaigns, identify trends, and recommend interventions. The project's ultimate goal is to equip Police Authorities (PAs) with the tools, knowledge, and training needed to effectively counter disinformation, enhance public understanding, and improve collaboration between police authorities and partner organizations.

2. Introduction

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 PAs by developing an integrated platform of advanced disinformation identification and analysis tools and technologies. It aims to employ state-of-the-art artificial intelligence (AI) methods while at the same time preventing any ethical or legal problems, for example, by using only aggregated data which cannot be attributed to individual social media users. The platform will cover disinformation from major sources (e.g., social media platforms and fake 'news' websites), in all modalities (text, image, video) and in multiple languages. Moreover, it will inform PAs of intervention approaches that have the most chance of success. By using the platform, PAs can be better prepared to investigate criminal activities linked to disinformation campaigns, and ultimately even prevent them.

The objectives of work package 5 (WP5) are to identify the social drivers and behavioural dynamics behind disinformation campaigns (WP5.1), and to develop tools to help PAs to intervene in ongoing disinformation campaigns and inoculate susceptible groups (WP5.2). This document aims to set out the final result of task 5.1 and provide insight and context to the *Impact Analysis Tool* that has been developed. Alongside the development of technical tools, this work also assists in the sensemaking of disinformation by means of PA training (WP7) in working with the VIGILANT system. An overview indicating the place of WP5 within the VIGILANT project can be found in Figure 1.

The rest of this chapter discusses the purpose of this document, generic concepts of disinformation that are relevant to the work described in this document and a description of EU projects also working on disinformation related problems that VIGILANT has cooperated with.

2.1. Overview

The purpose of this document is to describe the methods and the tools that have been developed for analysing the potential impact of disinformation campaigns as part of the VIGILANT platform. The various methods and elements used in the tool will be explained and applied to a use case of the Departament D'interior - Generalitat De Catalunya (INT) (Catalonian police) to demonstrate its practical relevance.

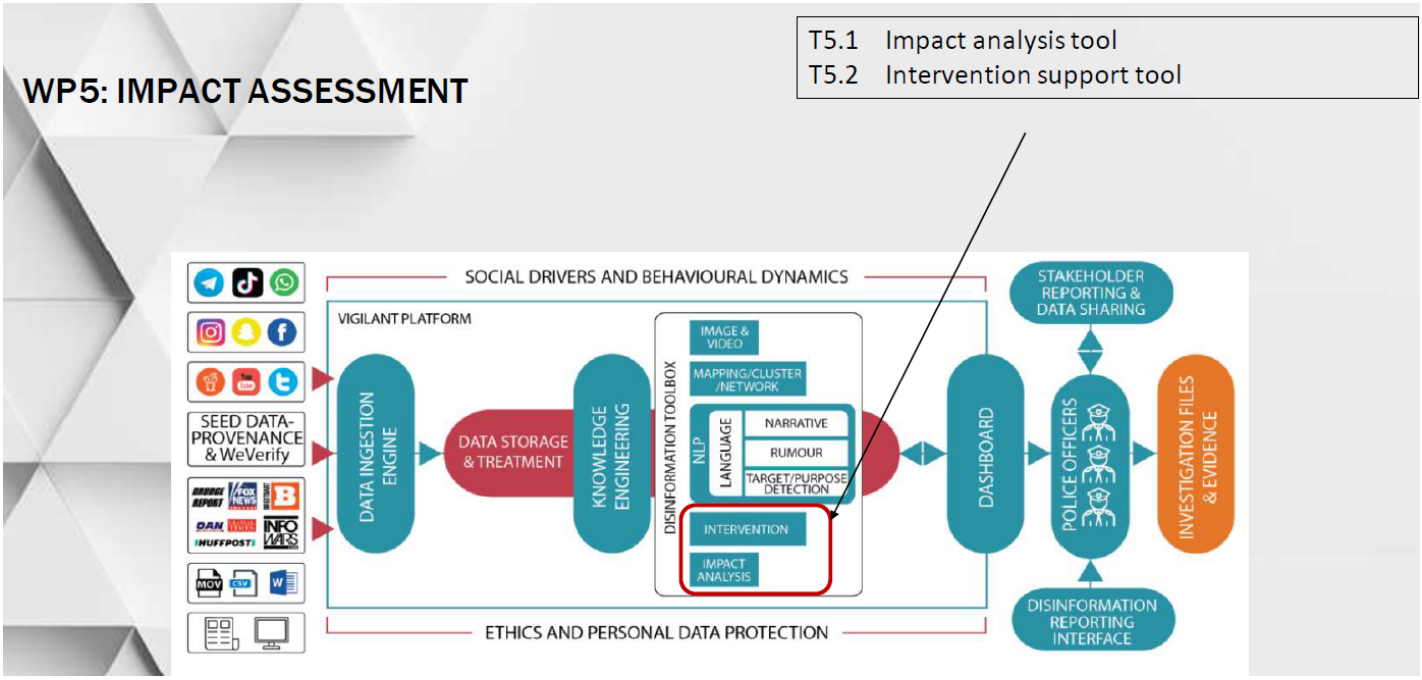


Figure 1: Overview of the work packages of the VIGILANT project. Highlighted is WP5, under which this deliverable falls.

This tool will draw on the analytical components developed in WP4. The content characteristics and profile types from these analytical components are structured in a database designed by WP3 and labelled according to the insights from the conceptual C5 model that was developed in T2.4. This combination of analysed content with structured conceptual insights will deliver an impact analysis tool that supports security professionals in understanding which disinformation campaigns are spread, why some resonate and spread more easily than others, and the likely impact this will have on society.

The work in this report builds upon previous work done with a thorough theoretical base and conceptual framework of disinformation from the Vigilant Deliverable D2.4 (Kruijver, Cadet, Finlayson, & Meer, 2023), which provided conceptual input for the technological tools of both WP4 and WP5.

The Impact Analysis Tool brings us a step closer to bridging the gap between academic research on disinformation from a conceptual view and the everyday reality of those combatting it with real world data, in organisations such as European PAs. To illustrate the workings of the tool it was applied to a real use case relevant for a partner PA – the spread of hate speech in Catalonia – which demonstrates the tool's practical applications. More use cases were consulted to help identify relevant functionalities of the tool and will also be used in later stages of the VIGILANT project to fit the PAs needs. PAs can use the impact analysis tool to assist in the sensemaking of disinformation, to structurally compare disinformation campaigns, or to help prepare a report for their supervisors and decide on possible counter measures. Moreover, since most European PAs have only recently created units to combat disinformation, a shared knowledge and understanding of the disinformation phenomenon is often still lacking.

Trainings on the theory (using the C5 Interaction Model) and the use of this impact analysis tool at a later stage of the VIGILANT project can contribute to the institutionalisation of necessary knowledge on how to identify and investigate disinformation.

VIGILANT is aware that other (EU) projects are also developing tools and methods to combat disinformation. The focus and/or domain of these projects is often different from VIGILANT, however finding collaboration through information sharing is frequently undertaken to learn from each other and avoid pitfalls that others have already identified.

2.2. Disinformation & the role of police authorities

Many different definitions of disinformation are being used by scientists and government organisations. Yet, the core ingredients are often similar, especially the notion that disinformation is *intentionally misleading*. In 2018, the High-Level Expert Group on Fake News and Online Disinformation of the European Commission published a definition which is widely accepted and often referred to (HLEG on Fake News and Online Disinformation, 2018). This report uses the term disinformation consistently as described in D2.4 (Kruijver, Cadet, Finlayson, & Meer, 2023):

> *Disinformation is false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit.*

2.2.1. Disinformation versus misinformation versus malinformation

A distinction can be made between disinformation, misinformation and malinformation. While disinformation is false, inaccurate, or misleading content that is made and shared intentionally, misinformation is also false, inaccurate, or misleading information, yet *not* made and shared intentionally. Malinformation is truthful information that is taken out of context and, like disinformation, aims to manipulate and harm. However, investigations into malinformation are outside the scope of the VIGILANT project.

People sharing misinformation are unaware that the information is false, inaccurate, or misleading. It's hard to establish the source of an original message online and even more difficult to establish the intention of the source. It is difficult to identify the transition from misinformation to disinformation. Is disinformation still disinformation if, for example, social media users are sharing it while believing that it is true, accurate information?

This problem also shows how difficult it is to prosecute makers and/or distributors of disinformation: their intention is hard to prove. Moreover, disinformation generally is not completely false, but often a mix between facts and fiction. A discussion what is truth and what is false is easily becoming a slippery slope in which often terms such as freedom of

expression versus (governmental) censorship are entering the debate. Who decides what is truth and what is not, especially when disinformation is often a mix of both?

2.2.2. Actors, aims and effects

A variety of actors may spread disinformation. They vary from individuals, for example dissatisfied anti-government citizens, to small or large groups, for example extremist or terrorist groups, and to state actors such as foreign security services. Professional disinformation campaigns by state actors are sometimes called 'Foreign Information Manipulation and Interference (FIMI)'. Moreover, the aims of spreading disinformation vary as well. Some disinformation is spread for economic gains (e.g. by creating disinformation to draw internet traffic to websites with online advertisements, or to influence stock trading), other disinformation aims at undermining society and its democratic institutions (e.g. political institutions, media, jurisdiction, or influencing elections) and regularly states use disinformation to influence diaspora in other countries as well. In the end, the effects of all these different kinds of disinformation are similar: unrest, distrust and polarisation in society. Disinformation regularly leads to illegal activities such as riots, violence, hate speech, etc (Holvoet, March 2022).

2.2.3. Role of police authorities in mis or disinformation

The spread of disinformation can have a range of harmful consequences, such as threatening our democracies, polarising debates, and putting the health, security and environment of EU citizens at risk. The content and spread of both disinformation can therefore have societal impacts. However, disinformation does not always require a role for PAs. Examples of this can be shifting norms and values or (other) people than could somehow feel offended by cursing or other types of strong language without any criminal act or security risk. Although this could lead to more serious situations, it does not mean unlawful actions are already happening and police actions should take place. If no harmful intentions are at play, such as is often the case for minors, no police action is required. Police will not monitor and intervene on a playground where kids are at play, unless they are called upon. Even seemingly harmful content, such as threatening language might not be perceived as such to those who receive it, as it can be part of their culture or norms and values to talk like this and no offense is taken. PAs do come in play when people do take offense or experience harm and report these actions to police. Figure 2 shows examples of behaviours that can be of increasing interest to police as they could be considered precursors to criminal behaviours or a danger to public order.

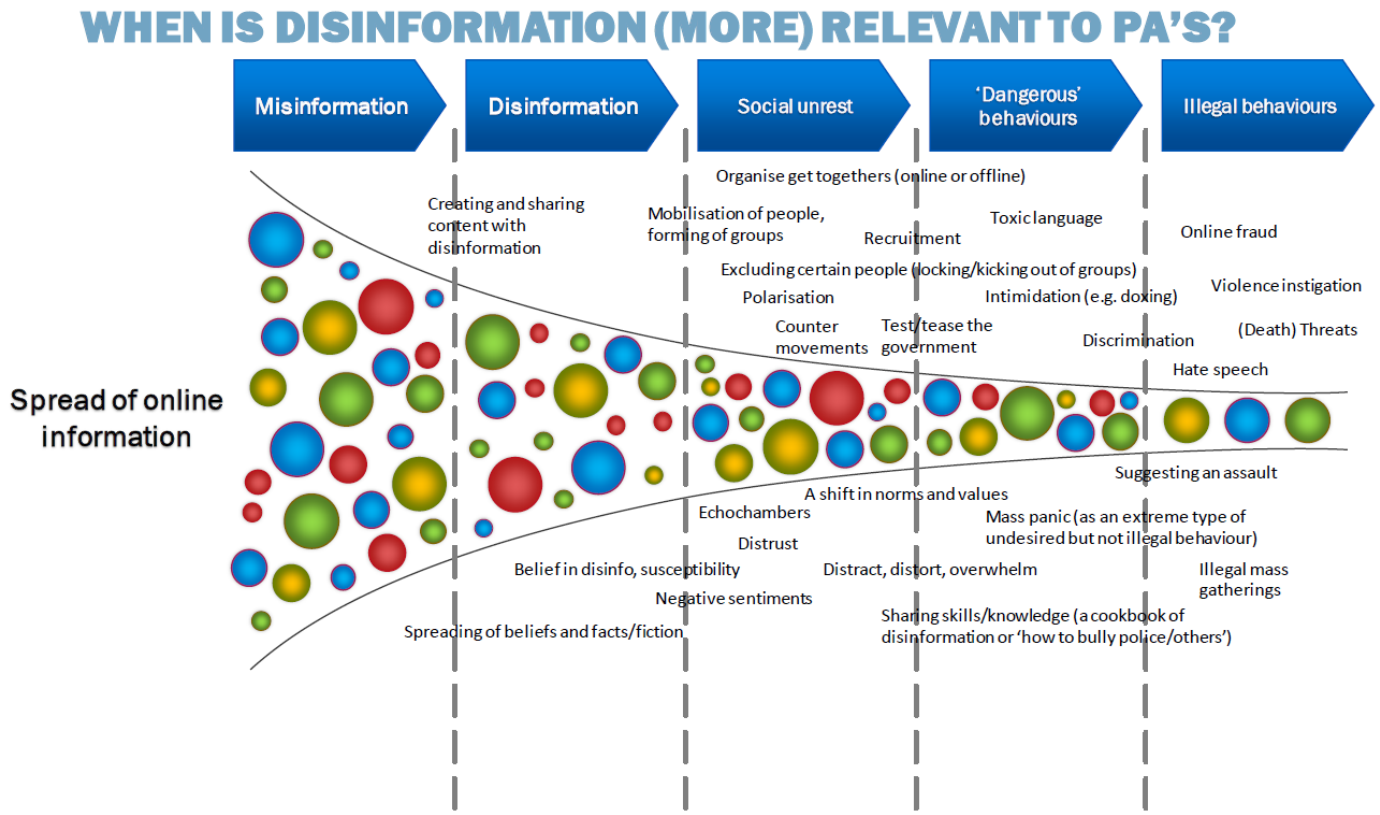


Figure 2: Possible stages of how mis- and disinformation becomes more relevant to PAs.

2.3. FERMI Project

The EU FERMI-project (FERMI, Fake News Risk Mitigator, 2024), a sister project to VIGILANT also looking into Disinformation and Fake News (D&FN), is developing a Community Resilience and Socioeconomic Watch. The project attempts to encapsulate FERMI's Community Resilient Management Modeler and the Socioeconomic Disinformation Watch into a joint component, to provide both insights on evolving digital threats and possible measures that can provide practical aid in tackling crime (FERMI Community Resilience and Socioeconomic Watch, 2024). It uses a Multi-Criteria Decision Analysis (MCDA), integrating more quantitative methodologies from ISO standards 31000 and 223XX, to provide a robust decision model capable of systematically evaluating risks stemming from an associated D&FN instance or crime investigation(s). The component follows a meticulous assessment whereby all high- or extremely high-impact indices signal that the system must provide countermeasures to the case under investigation. However, should the impact index be of a lower threshold, the system component will not advise any countermeasures. The indicators include socioeconomic factors, technological advancements and behavioural profiles.

Between the two projects knowledge has been exchanged on the approach and planned steps forward. One of the differences between the VIGILANT tool and the Socioeconomic Disinformation Watch is that FERMI follows a more

quantitative analysis and a fixed number of criteria in their approach with MCDA scores, whereas VIGILANT follows a more qualitative approach with a variety of indicators that can be used optionally where impacts are assessed that differ per use case.

2.4. DISARM foundation and framework

The DISARM Foundation is a non-partisan, non-profit organization that is incorporated in the United States but also works across Europe with several partner organisations (DISARM Foundation, 2024). The purpose of the DISARM Framework is to provide a 'common language' for the many people around the world working to mitigate the harms caused by disinformation, by enabling them to track, describe, and share their analyses using this common language.

DISARM is an open framework for those cooperating in the fight against disinformation and is in itself agnostic, not directing users onto any path. DISARM has been accepted as part of the official data exchange system on disinformation between the US Government and the European Union. The DISARM framework, designed for describing and understanding different parts of disinformation incidents, is based on the MITRE ATT&CK and STIX information security frameworks.

- DISARM-**STIX** for disinformation objects, including actors, behaviours, narratives and artifacts - this makes it easy for DISARM data to be passed between ISAOs and similar bodies using standards like TAXII.
- DISARM **Red** for disinformation creation behaviours (attacker's tactics and techniques)
- DISARM **Blue** for disinformation countermeasure and mitigation behaviours (defenders).

DISARM is potentially relevant in several areas for the solution VIGILANT proposes and possible pathways for a potential exploitation phase. On the data layer, the STIX protocol that DISARM uses (derived from the field of Cyber Threat Intelligence) is a potentially compatible solution for data exchange on disinformation. The DISARM framework is interesting for the exchange of knowledge on the different criteria to make sense of disinformation and the list of interventions that are part of the Red and Blue DISARM framework, which is still under continuous development. The VIGILANT tool can make use of these frameworks, as both the protocol and framework are open sourced. Besides these technological and knowledge sharing solutions, the DISARM foundation works with multiple communities of developers, researchers and practitioners. These are currently in other fields than policing, but the policing community in the EU, and potentially a wider global community, could join these. This allows for the development of a broader community of practice and connections with communities of scientists and developers that can assist in the scientific validation of criteria or interventions or provide new (open source) technology that could be used in policing practices.

2.5. Reading guide

The outline of this report is as follows: Chapter 3 explains the research methodology. Chapter 4 defines the concepts and models used in the impact analysis. Chapter 5 presents the tool, including a detailed description of its various elements, which is then applied to a use case in Chapter 6. Chapter 7 discusses some considerations for a potential follow up exploitation phase for VIGILANT. The report ends with a discussion of the work and possible future work in Chapter 8, which includes conclusions as well as limitations of this report, categorised along two lines: its academic contribution and its practical relevance.

3. Methodology

The impact analysis tool aids analytical tasks by linking online indicators of disinformation to possible effects that require or benefit from police intervention and providing a visual overview of these indicators. To achieve this goal, Task 5.1 followed multiple iterative steps to gain a structured understanding of the imprecise nature of real-world effects from online disinformation, as well as the technical, legal and ethical constraints involved in developing analysis tools for PAs.

This Work package first investigated the theoretical side of quantifying the usage and escalating behaviour of disinformation. In cooperation with WP2.4, possible indicators and effects were identified. These correspond to the inputs and outputs of the impact analysis tool. Conceptual models were established to model the behavioural flow from the indicators to the effects. These models incorporate principles from disinformation, ethics, and studies towards the modelled effects and system dynamics. Finaly, the impact analysis tool was developed with the goal for users to be able to create, analyse and maintain the models and ultimately integrate with the greater VIGILANT platform.

This chapter aims to explain key concepts to understanding disinformation and the process of developing the impact analysis tool. It begins with the exploration of system mapping, introducing the methodology used to model the complex systems of behaviour. Next, the process of identifying indicators and effects, which are relevant to detect and assess disinformation activities. Finally, the chapter addresses some of the ethical considerations that are involved with police activity of monitoring open-source platforms and disinformation. This section highlights the importance of balancing effectiveness with respect for privacy, abuse preventions and policing.

3.1. System mapping

Online and offline behaviour is difficult to follow, quantify, discuss and anticipate on a surface level. Many seen and unseen factors have an influence, ranging from the mood of an individual to the social-cultural norms of entire regions. It is therefore important to obtain an understanding of this complex system of behaviour that includes many of these factors to analyse the impact that disinformation has on a group.

The impact analysis tool uses system dynamics, specifically system mapping, to design models that represent the structure and behaviour of the online (and parts of the offline) environment. These models are made up of nodes (measuring the level or likelihoods of certain behaviours), directional connections (representing correlating behaviour), and feedback loops (circular chains of connections that amplify or reduce behaviours). A directed

connection between two nodes represents a correlation of the two behaviours over time (delayed effects, where a behaviour does not have an immediate effect, are still correlated). The source node of this connection is considered the indicator node, and the target node is the follow-up effect. When a certain behaviour can be measured using technical components or humanly assessed, the value of its node can be fixed. It therefore becomes an indicator of the entire system, as the current state of the system can be determined by the measured indicators. The effect of the model depends on the focus of the analysis and most often these refer to escalating behaviour for which police intervention is possible or required by law. A node on its own, without specifying it as a source or effect, is referred to as an impact type or as modelled behaviour. System dynamics allows for analysis and simulation of the interaction between different components of a system and how changes in one part of the system might ripple through and impact the entire system.

The largest benefits of system mapping also highlight some of its weaknesses. It faces challenges in capturing the complexity and non-linearity of real-world systems, which often involve numerous interacting components and feedback loops. Building a model requires balancing detail with comprehensibility. Overly simplistic models might miss critical insights, while overly complex models can become difficult to understand, analyse and manage. To manage the complexity, the system can be simplified to its most critical elements, focusing on the key feedback loops and interactions. If the resulting model is too simplified to be accurately used, additional details can be added to the model to increase the complexity. Accurately modelling these systems is also difficult, especially because crowd behaviour is an inherently chaotic system where small changes can lead to disproportionately large or unpredictable effects. Moreover, the availability and quality of online data is not guaranteed, which can impact the scope or quality of the analysis as well as the validation of the model. Fortunately, PAs often have expert analysts who can use historical data to fill gaps in the model. However, this can become labour intensive if done at a large scale. Additionally, the technical and measurable indicators are only a part of the puzzle. Some indicators cannot be directly measured, adding another layer of complexity when abstracting them, or adding unreliability by measuring them with indirect methods.

The system dynamic approach for the tool is designed to be adaptive. This means that when changes occur, such as new data and knowledge becoming available, models being proven incorrect, or changing laws and regulations, the tool can be adaptive. The iterative process helps ensure that models remain relevant and accurate over time, providing valuable insights even as the systems they represent continue to evolve.

3.2. Identifying indicators & effects

Effective monitoring of disinformation relies on accurate indicators that signal the presence and escalating characteristics related to false and misleading information. These indicators must be relevant to the specific disinformation threats being monitored and must be capable of adapting to evolving tactics. To ensure flexibility, the impact analysis tool uses a variety of indicators, either directly or indirectly measurable, within different models that can be changed and adapted to suit the needs of the user.

The identification of useful indicators starts by determining the scope of the current impact analysis. This includes specifying what the focus of the analysis is, such as what motivations and types of disinformation can be expected in this analysis. Here, the C5 model designed by WP2.4 can be used to ensure that all aspects of disinformation and escalation are considered and that each section is accounted for by at least one indicator when possible. It may not be possible to identify direct or indirect indicators for certain parts of the C5 model. These missing indicators become part of the analysis and 'indicate' missing situational awareness.

Direct indicators can be measured directly using automated tools to gather large volumes of data from various platforms. This often results in quantified data. WP4 has developed various tools and methods to quantify different indicators, such as image text recognition, sentiment analysis and hate speech detection.

Indirect indicators cannot be measured but they must be inferred from different sources of information. For example, the use of 'persuasive messages' cannot be measured directly as we have no tools designed for detecting it, it can however be inferred from the use of emotions in messages, the writing quality and the use of fake or misleading promises.

The validation of found indicators is a twofold process. First, the technical component of an indicator has to be tested on accuracy. During the development of indicators, certain performance metrics are identified and used to determine if the component behaves as it should, and accurately indicates the behaviour it is trying to represent. This type of validation is further detailed in the development documents of WP4 and the development of the individual components. The second validation step is determining if a component is suitable for the current model and whether the computed representation is correctly interpreted and used in the model. This occurs when the model is constructed. Experts provide arguments for the inclusion of the indicators in a new model. However, before new models are used by PAs, the models should be validated through tests, set use cases and an extensive exploitation phase. These evaluate all aspects of the model and examine in which situations the indicator and/or model fails to

accurately represent the current state or trend. Performing a full inspection of a new indicator or model is a lot of work, and it is difficult to standardise the evaluation process, as it depends on the type of model, the cases for which it is designed and local rules and regulations. A multi criteria validation analysis can be combined with system dynamics (such as used in FERMI) for full diagnostics, but this is not within the current focus of the VIGILANT project. But, if a validated policy measurement is to be developed, then this can be combined with system dynamics.

3.3. Ethical considerations

It is crucial to recognize that the analysis of possible future events, and the subsequent actions of PAs on those analysis, can quickly enter unethical grounds. Particularly when the tool assesses the current situation and trends to create an understanding of possible future situations, which is subsequently used to offer suitable interventions that the PA is able to apply. It is therefore imperative that the impact analysis is developed with ethical considerations in mind that translate into ethical safeguards when the tool is implemented. An early assessment of the ethical challenges of the impact analysis tool is described in document (Valentic, 2024). This assessment identifies ethical concerns relating to privacy, bias, accuracy of analysis, and decision-making transparency. To ensure an ethically sound tool, the assessment recommends a focus on human oversight, continued validation, documented transparency, active bias elimination, and clear accountability.

These recommendations were followed up during the development of the impact analysis tool to influence design choices and create an ethical analysis process. An important part of this is the use of a qualitative analysis where humans maintain control and all decision making. The aim of the impact analysis tool is to aid analysts in their work, not to take over or automate the entire process. The analyst is therefore able to provide adequate oversight whilst remaining accountable. The analyst is further encouraged to maintain good and alert decision making by checking questions that are added within the model to stimulate critical thinking. These check questions also provide clear points where biases in decision making can be considered and mitigated. The formalisation of the decision making of an analyst, together with the checked questions and scientific documentation of the models, provides an explicit method to document the factors leading to a decision, ensuring the process is transparent. Furthermore, the impact analysis tool is designed to be evaluated and updated continuously to ensure the relevance and accuracy of the models within complex dynamic environments.

The further implementation of these ethical measures is described when features and design choices are explained.

4. Models for Supporting Impact Assessments

4.1. Overview

The impact analysis tool employs a system mapping approach using multiple frameworks to analyse the impact of disinformation comprehensively.

The C5 behavioural interaction model (Section 4.2) focuses on how disinformation affects individual and group behaviours. It examines the interaction between different factors that influence the process of disinformation, from creation to dissemination and wider societal consequences. The C5 model helps identify patterns of disinformation interactions and how it influences behaviour and actions. Additionally, the escalation model (to be explained in Section 4.3.1) tracks how disinformation can intensify and spread over time. It focuses on illegal acts that require police intervention. It captures the dynamics of how disinformation escalates, influencing larger groups with broader effects and more intense actions.

Together, these models are used to create the effect model (Section 4.3). It synthesizes insights from both the C5 model and escalation model to analyse the overall impact of disinformation. While both the C5 and escalation models are quite abstract, the effect model uses measurable indicators to provide a current, interactive, and comprehensive view of how disinformation not only changes individual behaviours, but also escalates to affect societal structures and systems, offering valuable insights for designing effective interventions and mitigating strategies.

Although the effect model uses quantified indicators within its components, the model itself has a qualitative nature. The digital online ecosystem suits itself well to automatic data collection and processing since public sources of information are used that include some level of structured data formatting. However, the models of the impact analysis require a certain amount of human intuition that a fully quantified model doesn't allow. Effect models have to adapt to the analysis task and current situation and can therefore be quite different from each other. An example of an effect model for analysing 'mass gatherings' is shown in Section 4.3. But to provide distinction, a 'fraud' effect model has also been added to the appendix. The fraud model was made to study the differences between models for political and economical motives, the motives from the C5 model.

4.2. The C5 Interaction Model

Making sense of online disinformation, or other harmful online content, is challenging. PAs can benefit from better systems that help them to detect and analyse such information, not just from a technological standpoint, but also conceptually and theoretically. The VIGILANT platform has been designed to reflect a whole-of-society understanding of disinformation that is more than just technological tooling. As such, the platform is underpinned by a suitably comprehensive conceptual and theoretical framework which aggregates insights from various fields in the social and behavioural sciences, including communication science and media studies to anthropology and psychology. The C5 Interaction Model, developed by WP2.4, allows for greater understanding of the different factors at play in the creation, spread, and effects of online disinformation. The model forms the theoretical bedrock upon which the aim is to make it easier for end users, such as operational analysts in PAs, to detect disinformation by breaking down the various key elements that comprise disinformation but also to improve the analysis of disinformation by providing insights into the causes, content and consequences of any given piece of disinformation.

The model contains five main elements that play a role when exposure to disinformation content leads to cognitive and behavioural effects: Context (social, cultural, political or economic factors, important events or relevant trends), Causes (creators and their motives), Content (the tailored piece of disinformation), Consequences (short and long-term effects on the consumer and society), and the Cycle of Amplification (the interaction of receiver susceptibility, dissemination factors and the possible interventions to counter the cycle). A concise overview of the C5 model with the interplay of the C's can be found in Figure 3. The various elements in the model will be especially useful to support the development of the analysis and detection tool in WP4 and WP5 of the VIGILANT project, particularly the Impact analysis tool (WP5.1) and the Intervention support tool (WP5.2). Details on the methodology and development of the C5 Model can be found in deliverable D2.4 of the VIGILANT project (Kruijver, Cadet, Finlayson, & Meer, 2023).

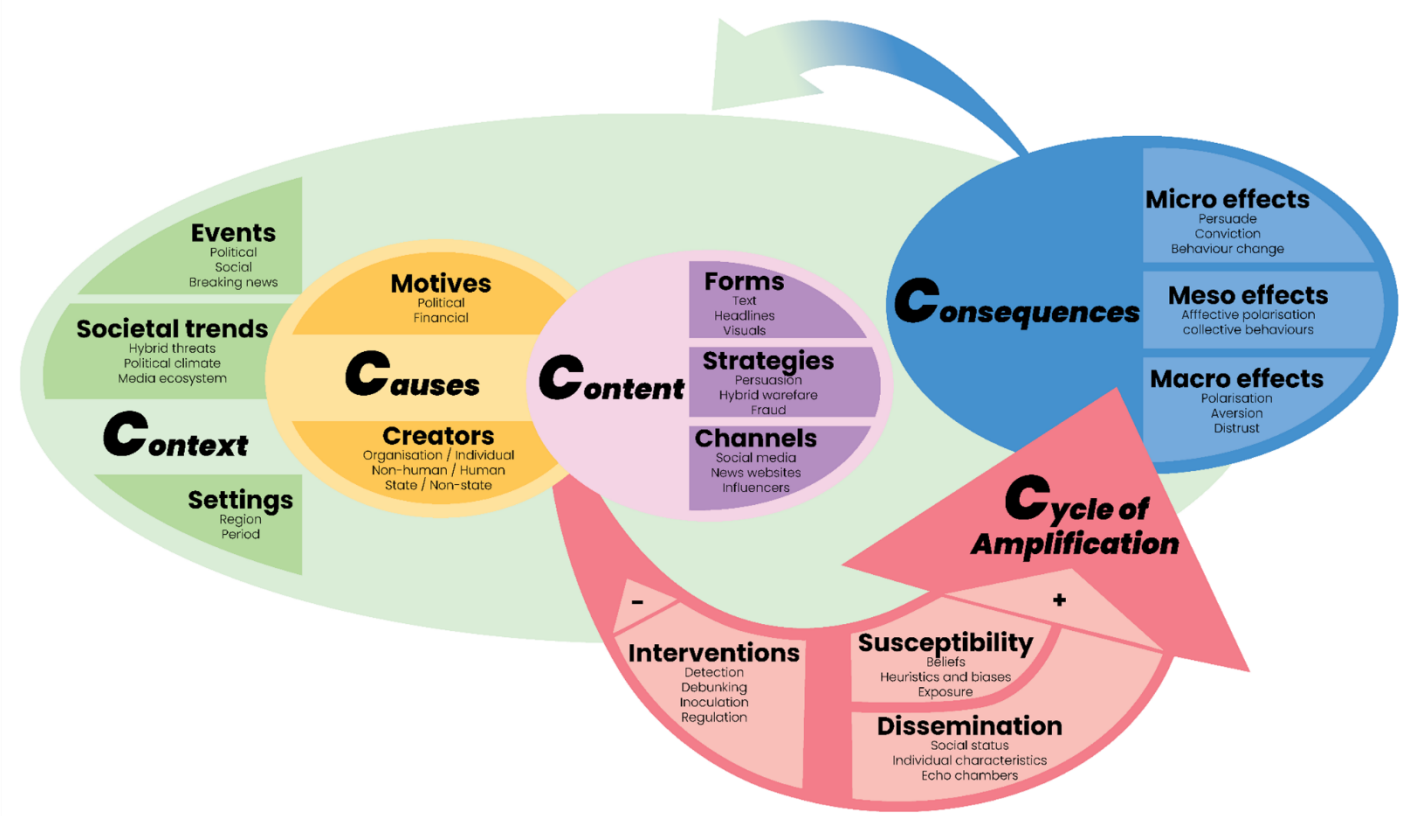


Figure 3: The C5 model which examines the interaction between different factors that influence the process of disinformatin, from creation to dissemination and wider societal conseqences. Source: TNO, 2023

4.2.1. Context

Disinformation is considered to be a 'context-bound phenomenon'; context is fundamental to understanding the actors, intentions, and techniques behind the manipulation (Hameleers, 2023). Context refers to anything related to the social, cultural, political, or economic setting or environment, including events or trends. This is arguably the most important factor to consider when analysing disinformation.

4.2.2. Causes

This element comprises two main sub-factors that help define the source of the content creation: the creators and their motives. Causes here refers both to 'root causes' – although this is more adequately covered by Context – as well as more immediate and short-term causes of disinformation. Causes interacts with Context as contextual factors may create opportunities for creators and their motives. In other words, creator's motives may involve the exploitation of a specific opportunity that arises from favourable circumstances, such as an event or social trend.

4.2.3. Content

Content relates to the constituent parts of any given piece of disinformation. For example, what form does the disinformation take (e.g., text-based or visual)?; how is it being disseminated?; what action or strategy is being propagated?; and which narratives and other storytelling-devices is the disinformation exploiting?

4.2.4. Consequences

Consequences relate to the factors that are the direct or indirect results of exposure to disinformation, especially visible in the behaviour of its recipients – in other words, the impact of disinformation. This can happen either at an individual level (micro effects), which then emerge at the group level (meso effects), or even escalate to the societal socio-psychological level (macro effects leading to public disorder). There is a hierarchical and escalatory relationship between the micro, meso and macro levels, which can result in detrimental offline consequences.

4.2.5. Cycle of amplification

The Cycle of Amplification refers to the relationship between dissemination, or propagation, and persuasion, which is usually the overarching goal behind disinformation (George et al., 2021). As such, this element demonstrates the interaction between the first four elements: within certain context (C1) disinformation messages are created (C2). Based on its content (C3), the susceptibility of its receivers (C5) and the dissemination (C5), effects are created (C4). This in turn changes the context and can become causes in itself for new disinformation campaigns. To counter this amplifying effect, PAs can employ interventions (C5) to mitigate or prevent the spread of disinformation.

4.2.6. C5 model summary

The C5 Interaction Model brings us a step closer to bridging the gap between academic research on disinformation and the everyday reality of PAs and relevant practitioners. The C5 Interaction Model provides a conceptual foundation for further endeavours of the VIGILANT project to train police officers in better apprehending disinformation processes. The research demonstrated that the context is of the utmost importance when it comes to understanding and thereby mitigating the spread and negative effects of disinformation. By using the C5 Interaction Model to analyse what factors are at play, one can better understand the interaction between certain events, for example, and the creation and dissemination of disinformation.

This is the key contribution of the C5 Interaction Model – a focus on analysing the interaction between different factors that influence the process of disinformation, from creation to dissemination, and wider societal consequences. Whereas communication models have drawn up the communication process and other publications have identified

the Cycle of Amplification, this research connects all the factors, whilst demonstrating their fluidity. The comprehensive C5 Interaction Model shows that there is not one fixed route that a piece of disinformation takes, making it a highly complex phenomenon to research. This conceptual understanding of disinformation can enable PAs to anticipate disinformation and mitigate its effects on society. Figure 4 below gives a concise overview of the "C3 Content" part of the C5 model and the technical components that are mapped to indicate which parts they can represent. The experts that design impact analysis models can ensure their model touches all disinformation aspects, by fully mapping each part of the C5 model with indicator components that are used for the model. This method can also highlight gaps in understanding and a lack of measurable information.

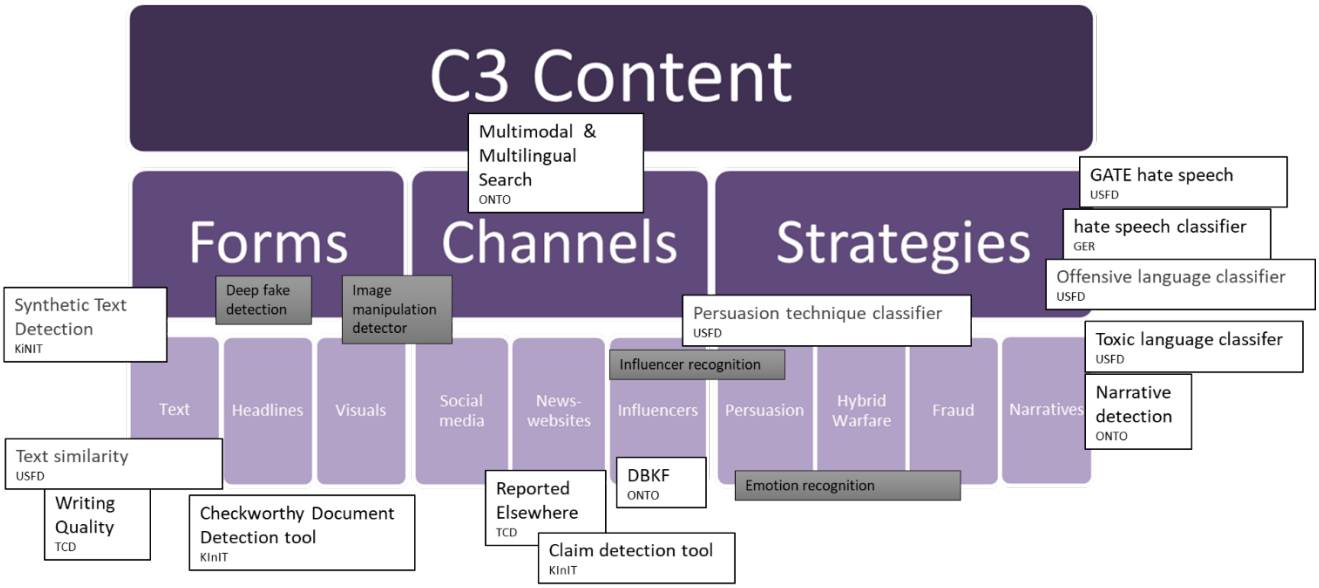


Figure 4: Mapping of VIGILANT analytical components to "C3 Content" elements of the C5 model. Grey items represent components not (yet) developed in the VIGILANT project.

It should be noted that this is not an exhaustive model. The environment in which online disinformation takes place, which also interacts with the physical world, is constantly changing. Therefore, the factors that were identified can best be seen as thematic labels under which numerous specific examples can be categorised. Nonetheless, the C5 Interaction Model can be used to create an understanding of disinformation messages and campaigns, including their effects. By applying it to practical examples, the model can help to analyse what different factors are at play. This can be helpful in identifying potentially dangerous interactions between factors.

4.3. Indicator based models and escalation model

Future-oriented analysis can be divided into three groups: trend analysis, future explorations and early detection. More forward-looking analysis (such as forecasting instead of backcasting) allows for more proactive anticipation of important developments and events in the future. With indicator-based risk analysis (IBRA) a more proactive approach for the third form of future-oriented analysis of early detection is possible, to prevent larger, undesirable impacts on society. IBRA contributes in particular to the early detection of undesirable developments in public safety and security problems.

IBRA is a group of analysis methods that considers events or trends as part of a bigger societal problem in which there is interest, and which potentially poses a threat (Keijser, Wessels, & Vries, 2023). IBRA is applied for early detection of changes in risks, such as for example changes in behavioural indicators. It is possible to apply IBRA to various public safety and security problems, such as the potential rise of protest movements due to social unrest in which disinformation plays a role, early detection of fraud in which disinformation plays a role, or the domain of surveillance and personal security in which disinformation leads to threats to people that are under police protection.

The application of IBRA has added value for the analyst (Keijser, Wessels, & Vries, 2023). Firstly, these methods ensure that the analysts' work is more model-based and can therefore be carried out in a more structured manner. Secondly, risk assessments are now much more widely based on available data and information and thus the traceability to underlying data and information is increased. Thirdly, these methods allow analysts to think process-wise about future developments and possible scenarios. Lastly, depending on the size of the model, the task of the analyst is placed in a bigger picture of the wider problem, enhancing sense-making and coordination between different policing tasks and partners of the police. This supports a better structured professional judgment (Keijser, Wessels, & Vries, 2023).

4.3.1. Escalation model in the case of mass gatherings

Social unrest is grounded in psychological mechanisms and looking at it from a psychological perspective can help to describe this process. Psychological modes can make sense of situations by analysing the exchange of individual (mental) states, contagion of the public space, and synchronisation to a collective state. As well as framing the specific events, the orientation of them and actions that can emerge when orchestrating or being influenced by an activist movement (Bar-Tal, Halperin, & Rivera, 2007; Drury, Reicher, & Stott, 2003).
There are five conditions for individuals to associate themselves to an activist movement, from a psychological perspective (Stekelenburg & Klandermans, 2013):

1. A grievance stemming from relative deprivation, frustration, or perceived injustice

2. It is perceived as possible and effective to alter the situation through protest
3. There is a political or politicised collective identity
4. There is a strong group emotion, usually anger
5. People in the group are connected in their grievances and encouraged to act (social embeddedness)

An in-depth analysis of conspiracy theories after the murder on Pim Fortuyn in the Netherlands showed that not all conditions were met between 2000 and 2014 (Buuren, 2013; Buuren, 2016), leading to escalations with real-world consequences. Politically driven social unrest emerges in a societal process of various steps (Renn, Jovanovic, & Schröter, 2011). We describe the pathway of emergent social unrest over time in terms of intensity and degree. Both escalation and de-escalation take place over time. The analyst can make these judgements in the VIGILANT system.

The intensity of social unrest is described by types of actions (concurrently) taking place:

1. Communication of dissatisfaction (based on a grievance)
2. Organisation of protest groups
3. Mobilisation to acts of unconventional protest or violent acts
4. (Suggested-) Actions of violence or illegal behaviours

These stages (see Figure 5 below) can be linked to the number of participants, relative to the total population, exhibiting the various actions to establish an intensity of social unrest (see Figure 6).

Figure 5: Four stages of escalation in social unrest leading to mass gatherings.

Three elements of initial mobilisation of o acts of unconventional protest or violent acts need to be investigated:

1. **Grievances** are increased by events that take place and perceived ineffective response to these events, (perceived) inequalities, and discrimination could all play a role in fuelling grievances and emotions (Kapiriri & Ross, 2020; Taylor, 2019; Alsan, Westerhaus, Herce, Nakaschima, & Farmer, 2011).
2. Pre-existing polarisation and efficiency of government in dealing with these problems are important determinants of (political) **opportunity** for social unrest (Gelfand, et al., 2020; Bavel, et al., 2020).
3. In terms of **mobilisation** various groups may be the core of a protest movement. Populist and minority groups may for example be important. Fear, anxiety and helplessness are more passive emotions and these in turn could mobilise other groups.

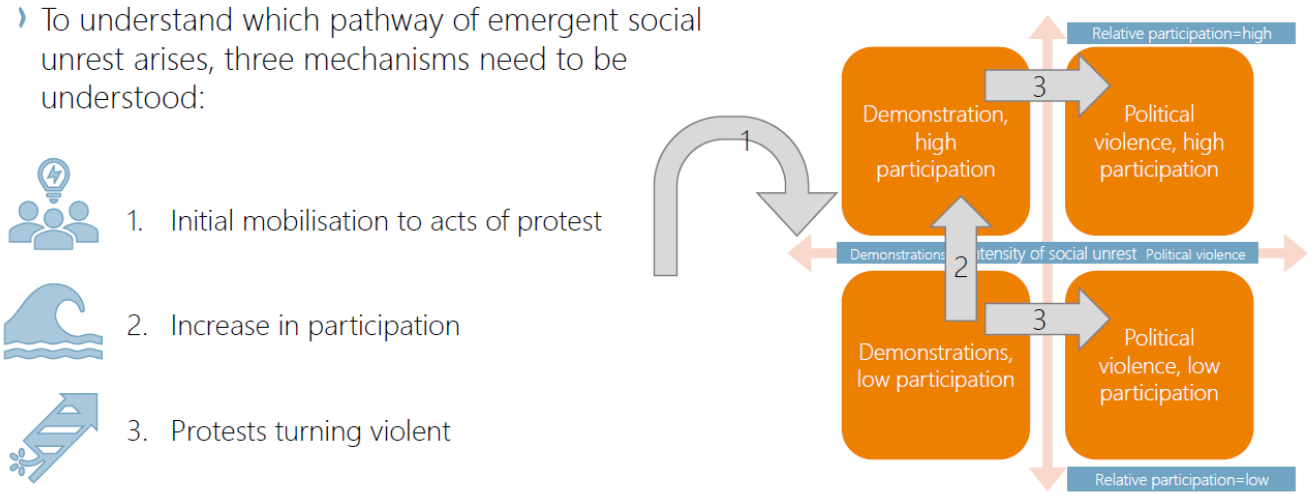


Figure 6: Pathway of emergent social unrest with impact indicators.

Larger groups of people will join the demonstrations when they already experience the grievance communicated in a starting protest movement. A specific large population segment might already experience a significant grievance, that is picked up by a protest movement (e.g. the Black Lives Matter-movement). Grievances or ideologies tap into grievances that are spread adequately among this population segment and this protest movement are perceived to be effective by a larger group of people. Economic hardship, psychological stresses, and lack of a good outbreak response that reduces economic/social damage can also play a role. Although this mechanism has not been studied in detail, some remarks can be made:

- Violent riots are more likely if there is distrust in authorities (Sullivan, 2019).
- Trigger events play a crucial role in protests turning violent. Reaction by law enforcement to demonstrations is an important determinant of demonstrations turning violent (Klein, 2012; Thomas, 2020).
- Interactions between various protest movements that act on diverging/contrasting dissatisfactions play a role in intensifying and demonstrations turning violent (Shiffman, 2020).

- Lack of changes in policy/government response to initial demonstrations or new measures and change in a societal problem increase a sense of hopelessness of protesters.

4.3.2. Mapping effect model impact types to real-world disinformation 'knot'

Narratives on a possible economic crisis on the horizon or already taking place, together with possible consequences such as upcoming government budget cuts and a decline in public services such as in education, might be messages clustered in narrative analyses, as different categories of fake news or at least speculations. Amidst these circumstances, a political event could be planned by politicians to counter these narratives and calm the public at large, but within a narrative cluster on this topic, false claims surrounding this political event might also be detected. These clusters of narratives could ultimately play a role and ignite the further creation and spread of disinformation. For example, a Telegram group that relates to such a situation could grow in both size and group activity, and together with a shift in sentiments as discussions on these topics, get more heated up. Furthermore, as they gain increased attention, from both inside as outside the group, more users join these, and similar, groups and discussions. This leads to the creation and spread of more disinformation.

As these narratives escalate both in size and sentiments, this could in turn lead to possible instigations. These instigations could point to people, locations or events and manifest either online or also offline to help mobilise people to make a difference, for example through a mass gathering such as a protest. In this mobilization online or 'mobbing', individuals could also decide to go a step further and express more extremist behaviours or views or even plan a personal assault or an attack at an event or location.

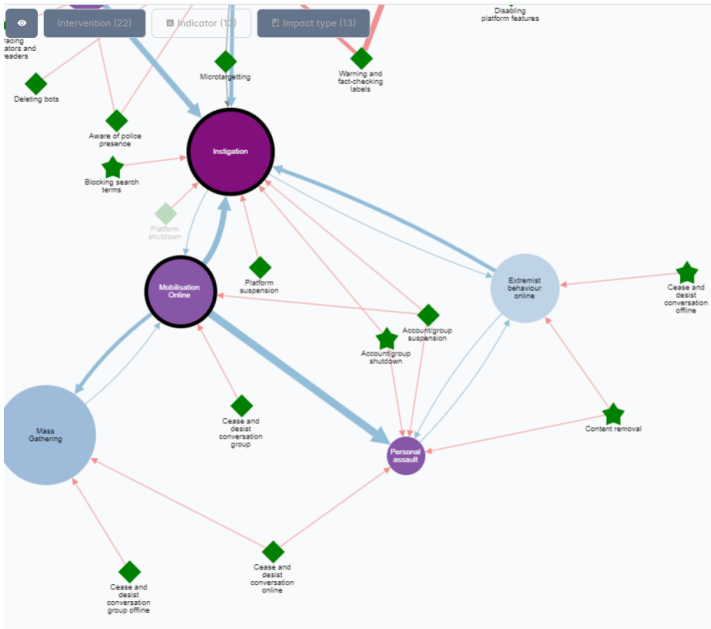
4.3.3. Mapping (technical) indicators per impact type of the mass gathering model

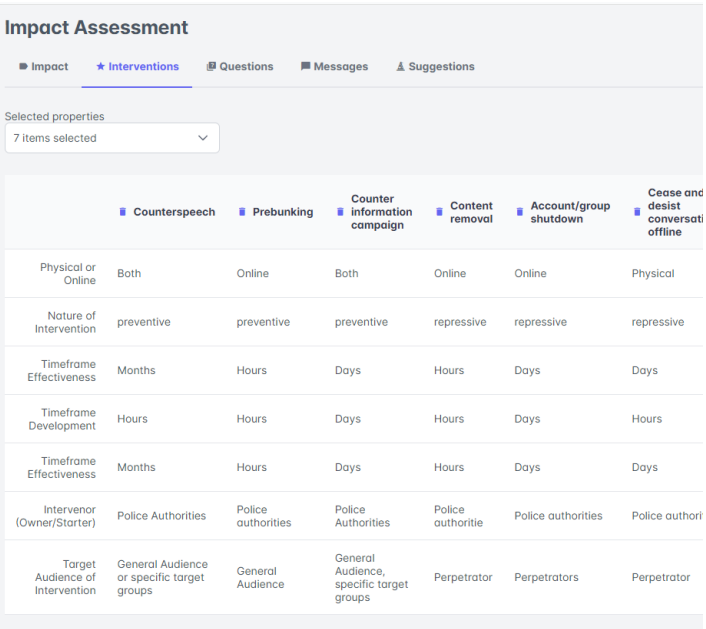
Narrative analysis: narrative clusters such as economical crisis, government budget cuts, decline in public services, political event.
Stance: towards narratives, towards disinformation (such as DBKF), towards events and hate speech
DBKF (Database of Known Fakes): as indicator for the spread of disinformation, both measuring the occurrence and amount of (similar) fakes.
Writing Quality as indicator for disinformation.
Emotions: Fear, Surprise, Disgust, Anger towards narratives, as an indicator of group sentiments towards disinformation, towards instigation, towards locations, events or people (personal assault).
Hate speech in instigation (as hate in instigation events, locations, people or organisations) and as indicator for extremist behaviour (in general hate messages towards locations, events, people)

Event detection as indicator in instigation (locations, events or people or organisations) and as indicator in personal assault (in general emotional or hate-like messages as a detector for locations, events, people).

4.3.4. Mapping interventions per impact type

Instead of listing all interventions that are developed in WP5.2 and described in D5.2 here, listed below is a selection of interventions that could be relevant in case of a mass gathering. Figure 7 also gives an impression of how these interventions are depicted in the impact analysis tool.





	Counterspeech	Prebunking	Counter information campaign	Content removal	Account/group shutdown	Cease and desist conversati offline
Physical or Online	Both	Online	Both	Online	Online	Physical
Nature of Intervention	preventive	preventive	preventive	repressive	repressive	repressive
Timeframe Effectiveness	Months	Hours	Days	Hours	Days	Days
Timeframe Development	Hours	Hours	Days	Hours	Days	Hours
Timeframe Effectiveness	Months	Hours	Days	Hours	Days	Days
Intervenor (Owner/Starter)	Police Authorities	Police authorities	Police Authorities	Police authoritie	Police authorities	Police authori
Target Audience of Intervention	General Audience or specific target groups	General Audience	General Audience, specific target groups	Perpetrator	Perpetrators	Perpetrator

Figure 7: Interventions selected (started left) for a possible mass gathering and listed in a comparison table (on the right).

- **Counter speech:** Countering offensive speech by calling out the speaker towards a more socially acceptable form of speech.
- **Pre-bunking:** Whereas debunking relies on proving certain claims, narratives or stated 'facts' are not true, pre-bunking aims to prevent false narratives from taking root in the first place (Roozenbeek & Van der Linden, 2019).
- **Counter Information Campaign:** Leveraging social information to encourage people not to believe, endorse, or share misinformation.
- **Content Removal:** Removing content that is labelled as disinformation or potentially dangerous.
- **Account/Group shutdown:** Shutting down an account or group that is spreading disinformation. This often is a legal request from PAs to online platforms or service providers.
- **Cease and desist conversation:** A (warning) conversation as a preventive measure in order to try and reverse their path, which is currently heading towards more extremist and dangerous or criminal behaviour.

4.4. Interventions

Countering disinformation can be done in different ways. Actions or processes of interventions of PAs can range from communication efforts to counter narratives all the way up to arresting suspects that are considered to be part in criminal behaviours related to disinformation. Because PAs also have a vast network of public and private partners for various policing tasks, the VIGILANT project has tried to include a broader view on policing interventions. Policing in itself can be considered a task not executed by law enforcement alone, to quote Sir Robert Peel: “The police is the public and the public is the police”, also public and private partners can take part in countering disinformation.

Examples include private social networking platforms that enforce their own platform policies by removing accounts or content. Customers of those platforms could report to these platforms but could also report to police or other government agencies or NGOs. Without describing these networks and processes, a wider view of interventions is needed. PAs can ask these public and private partners to join in a collaborated effort initiated by PAs. They could coordinate efforts on phenomena, or without coordination, other partners could intervene separately from their own respective roles. The full list of interventions is described in WP5.2 and includes both online and offline interventions, both repressive interventions (short term) and preventative interventions (long term). Lastly, these interventions could be unique for PAs, but could also be about interventions done by public or private partner organisations such as online platforms or social workers.

5. Impact Analysis Tool

This section of the report aims to provide a comprehensive overview of the functionality and implementation of the impact analysis tool. The content is structured to guide readers through the tool's conceptual framework, its practical application, and its technical implementation.

5.1. Intended users

The tool was developed with two types of PA end users in mind. First is the senior analyst or academic that wants to capture their knowledge on the phenomena of disinformation in a systematic way in the tool. They might not be involved in day-to-day operations but have relevant insights either by experience, studying the topic of disinformation, or both. The tool will benefit most when multiple fields of expertise come together and combine their shared knowledge to create a more thorough understanding of the disinformation phenomena related to the tasks of PAs.

The second type of user, the operational analyst, will use the tool with a more practical objective, inspired and assisted by the background knowledge that has already been recorded by the first group of end users or by other peer users. It can range from making observations and assessments, investigating potential short- or long-term effects, identifying relevant indicators to monitor and analysing or investigating possible interventions.

5.2. Data model and knowledge base

At the core of the impact analysis tool is its data model, which captures the key concepts essential for its operation. This model is used for both input fields and visualizations, so it requires careful attention. The data model is structured as a (knowledge) graph, made up of nodes and edges. This setup helps establish relationships across the tool, enabling a thorough and connected assessment framework for disinformation. The graph structure allows for complex relationships between various elements. For example, an impact model for disinformation can be linked to several (societal) effects, which can then connect to different (technical and/or manual) indicators. The graph-based data model offers high flexibility, making it easy to add new concepts or relationships as the tool evolves. It is a natural fit with the networked structure of system mapping (as explained in Section 3.1).

Key concepts in the data model include: Impact types, Effects, Indicators, Interventions, Questions and Impact Models. Each concept within the data model is characterized by a set of potential properties that define individual instances. These properties and their interrelations are defined in a dynamic ontology, that can evolve over time. The ontology contains basic properties such as name, description, and creation timestamp, as well as concept-specific attributes.

Among the key concepts, interventions possess the most extensive and diverse set of properties. Interventions are characterized by a set of properties to describe the different dimensions of their implementation and effects. These properties include the intervention's modality (physical or online) and scalability; its targeted outcomes and associated risks; classification details such as escalation level; the nature and group of the intervention; temporal aspects covering development, execution, and effectiveness timeframes; stakeholder information including the primary intervenor, mandatory partners, and potential collaborators; the target audience. A more detailed explanation of the interventions is part of the work of WP5.2

5.3. Tool usage

Using the impact analysis tool involves two different modes: input and analysis.

5.3.1. Input

During the input phase, various concepts can be entered. This includes describing different impact models, effects, indicators, interventions, and questions. Based on the properties of these different concepts, a detailed and coherent assessment can be built.

- **Impact type**: An impact type is a potential impact. Each impact type can be specified in relation to the impact models they are part of.
- **Effects**: Effects are the relations between impact types or interventions and impact types. They describe how strong relations are and across what timeframes effects between relations can be seen.
- **Indicators**: Indicators are entered to describe specific measurements or signals of impact types. Indicators can be the output of components developed during or after the VIGILANT project by partners.
- **Interventions**: Users can input interventions that are applied to achieve or influence certain effects.
- **Questions**: Questions can be added to further investigate certain aspects of the impact models, such as effects and indicators. Questions help analysists in tackling bias or other potential legal, ethical, technical or procedural issues that might right attention.
- **Impact Models**: A graph consisting of impact types, effects, indicators, interventions and check questions. Designed to analyse trends towards criminal activities. Users can enter various impact models that form the foundation for further analyses.

Each concept has its own input screens for adding and editing data. Although there are no built-in limitations, this part of the tool is intended to be used by the senior analyst to provide the background knowledge that will be needed later in the analysis phase.

In Figure 8 there is an example of an effects input screen. On the left, all potential effects are listed, selecting any of them will open a detailed input screen on the right. These will include general input fields like name and description, but also classifications like the escalation level and C5 area where each effect is active (as described in Chapter 4).

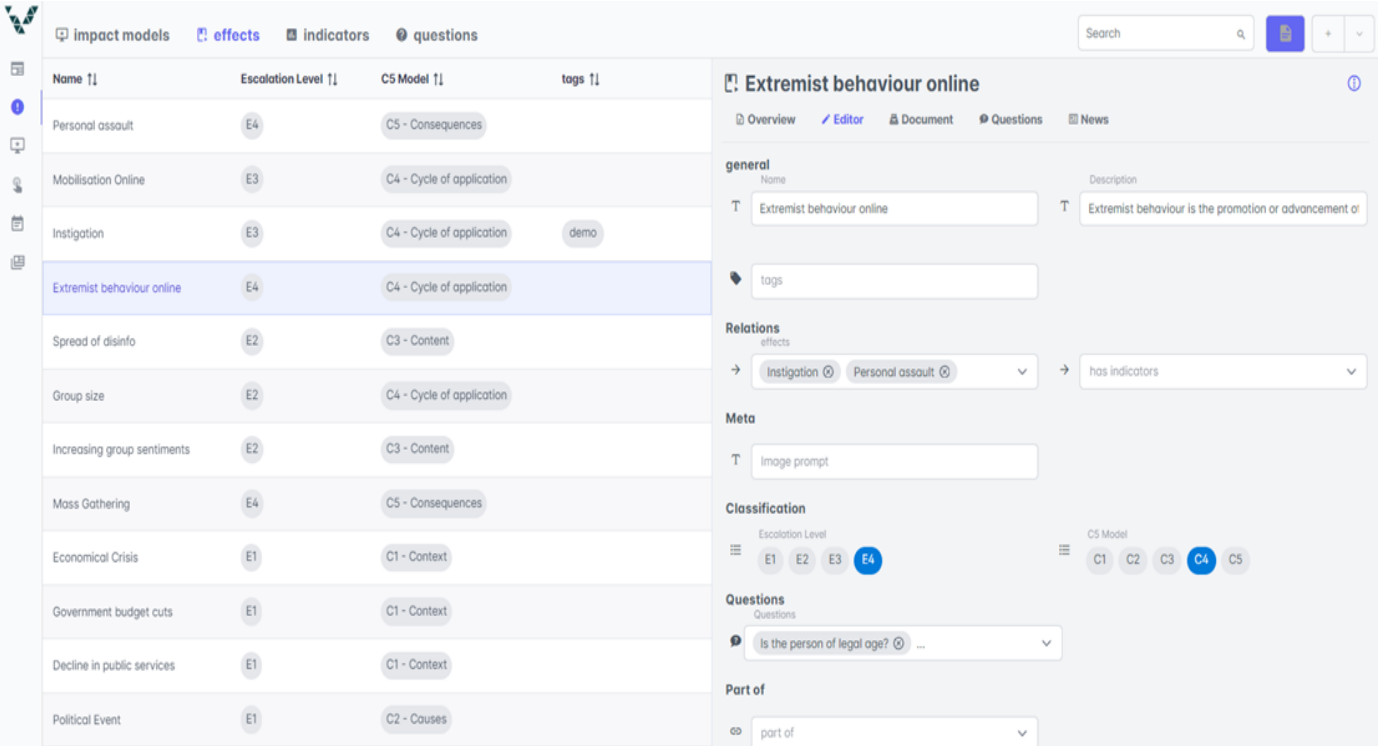
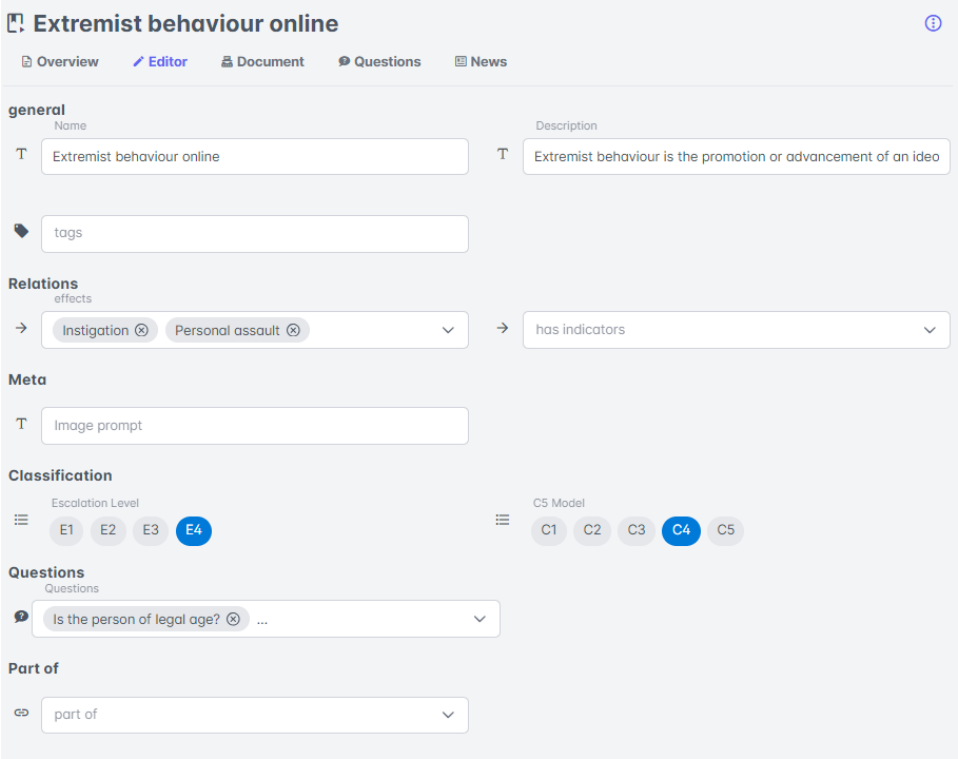


Figure 8: Effects overview and editor screen, below is a close-up of the editor for clarity.



Any relations to other concepts, such as potential other effects and questions, can also be selected here. For linking questions to indicators, interventions, and effects, a special tab page was created where a long list of questions is presented for each indicator, intervention or effect a long list of questions is presented. Analysts can simply select the most relevant questions that apply to their case (See Figure 9).

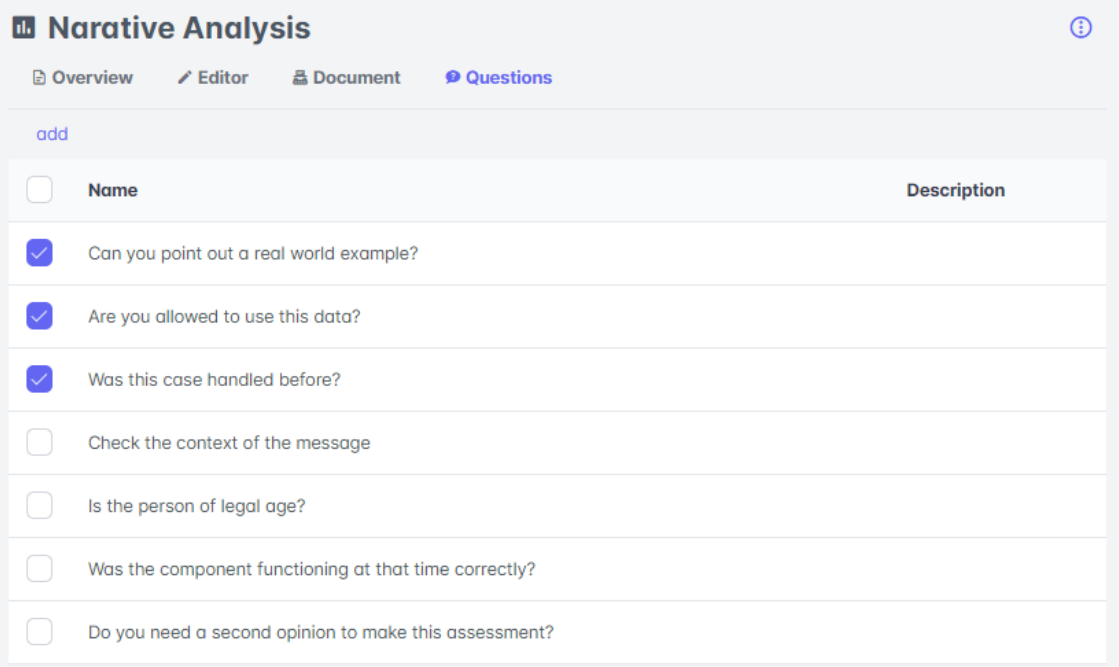


Figure 9: Adding check questions to indicators, interventions or effects.

5.3.2. Properties and descriptions

In addition to the various properties of each concept, the tool also offers the ability to provide a detailed description. This helps to clarify and contextualize the entered data, allowing for a deeper understanding and more accurate analysis, as shown in Figure 10.

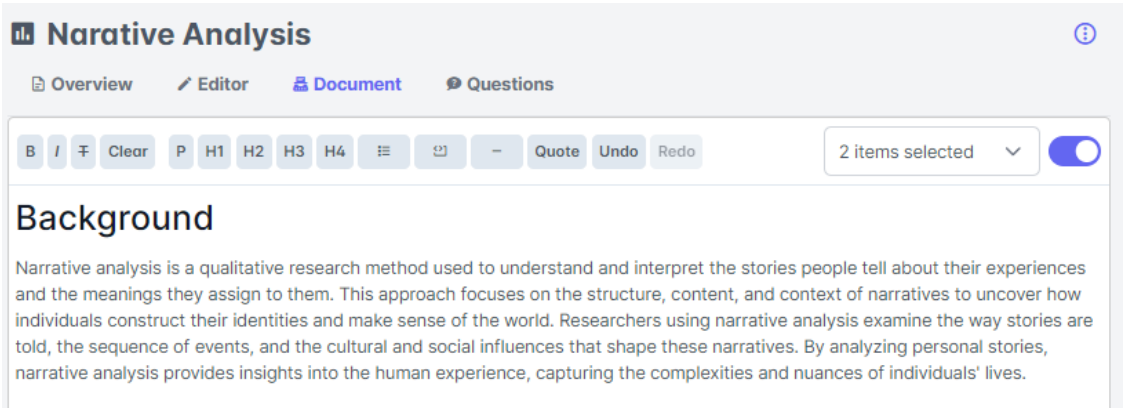


Figure 10: Description editor to add context or clarifications to each concept.

Through this structured approach in the input phase, users can enter relevant data, which enhances the quality of the subsequent analyses. The tool thus facilitates a systematic and thorough approach to impact assessment, from input to visualization and interpretation.

5.3.3. Filtering and searching

Some concepts can cause the list to grow substantially. With multiple labelled columns in the table, filtering and search functionalities are essential to navigate the list effectively. The dashboard offers several filtering options on the left side, depending on the active dashboard. For the list of interventions, it will include: Nature of Intervention, Escalation Level, Timeframe Effectiveness, and Category. Users can select one or multiple filters to narrow down the list of interventions. For example, filtering with a “preventive” Nature of Intervention and an Escalation Level of “E1” will display only those interventions that match both criteria.

5.3.4. Column sorting

Each column in the table, as seen in Figure 11, can be sorted by clicking on the column header. This allows users to order the interventions alphabetically by name, or sort by Nature of Intervention, Escalation Level, or Timeframe Effectiveness. This sorting feature is useful for quickly finding specific interventions or comparing them based on their attributes.

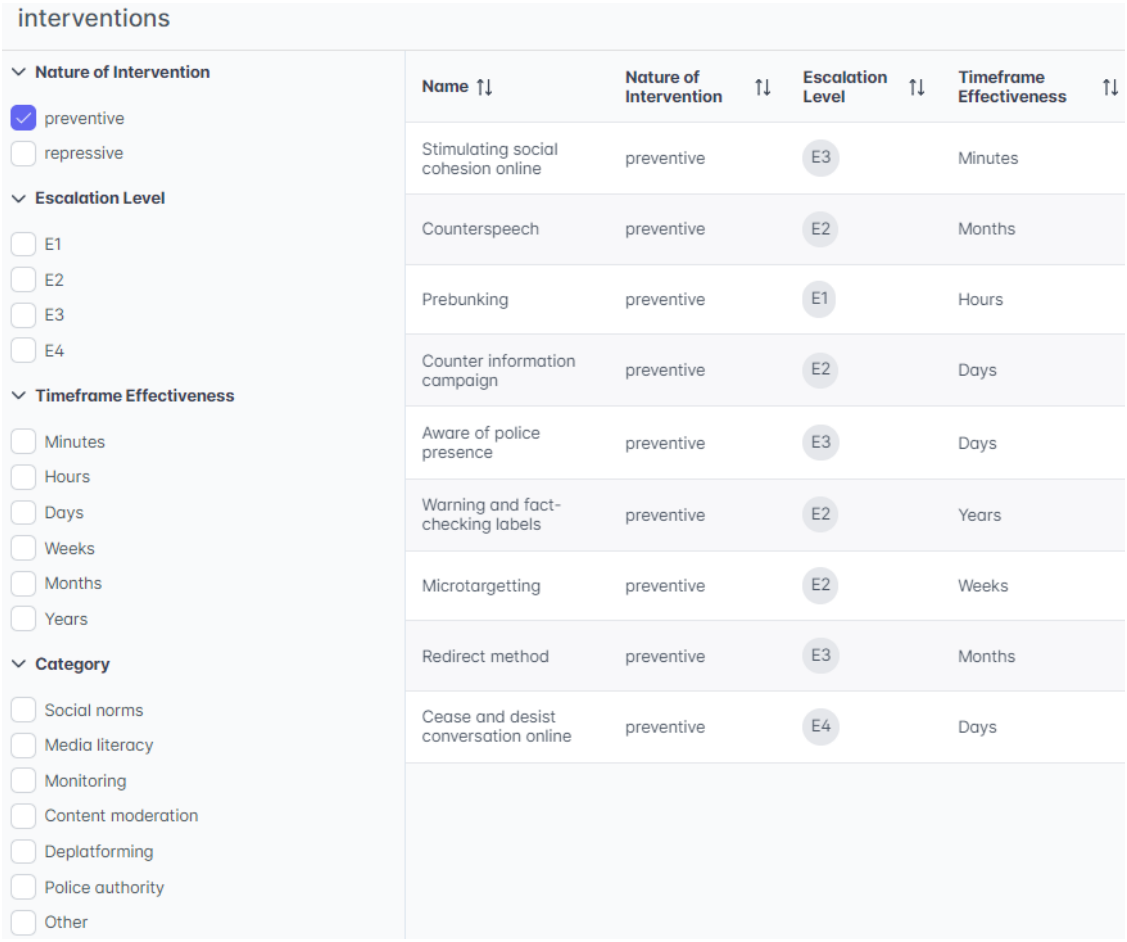


Figure 11: List filtering by various criteria.

The search bar enables users to find specific interventions quickly by typing in keywords related to the intervention name or other attributes. This feature is particularly useful when dealing with many entries, as it allows users to bypass the need to scroll through the entire list.

5.4. Impact analysis

The impact analysis consists of building and exploring system maps to establish relationships between effects, indicators, and interventions. The starting point for the impact analysis is selecting (Figure 12) or creating a new impact model. An impact model is a system map that is created by the users with effects, indicators and interventions available in the knowledge base.

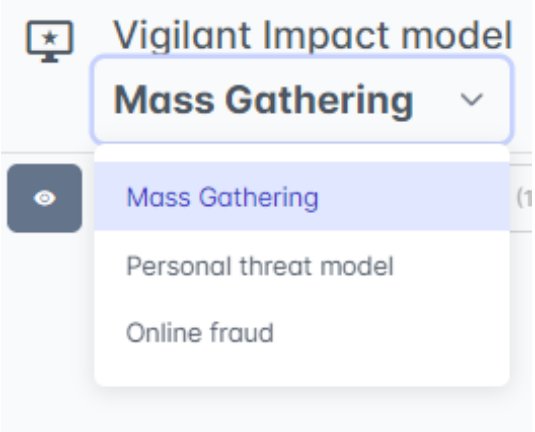


Figure 12: Impact model selection menu.

5.4.1. System mapping

Users can add these concepts to their map, create a visual layout manually and describe their relations. The main concepts of the map are the impact types and the effects describing how they related. Analysts can use these to identify potential escalating scenarios, explore indicators and manually assess the situation. It is also a starting point to investigate the effects of potential interventions or do a more detailed investigation of the messages that have been flagged on the platforms using the greater VIGILANT platform as integrated by WP6.

All concepts within the impact analysis tool can be reused across multiple maps, enhancing consistency and efficiency in the assessment process. However, the assessments associated with these concepts are unique to each situation. After defining impact types, an analyst can assess two key properties:

1. Impact Level: This is evaluated on a scale ranging from *very low* to *very high*, indicating the intensity or significance of the impact. The colour of the node is used to represent the impact level. It ranges from light blue for *very low*-impact assessments to dark purple for *very high*-impact assessments (The selection menu for the impact level is shown in Figure 13).

2. Scale: This property is assessed from *very small* to *very large*, representing the extent or scope of the impact. Larger nodes on the map represent impacts with a broader scope, while smaller spheres indicate more localized effects.

The combination allows for a nuanced assessment that considers both the intensity and the breadth of an impact. For instance, an impact might be rated as 'high' in level but 'small' in scale, indicating a significant but localized effect. Conversely, an impact could be 'low' in level but 'very large' in scale, suggesting a widespread but mild influence.

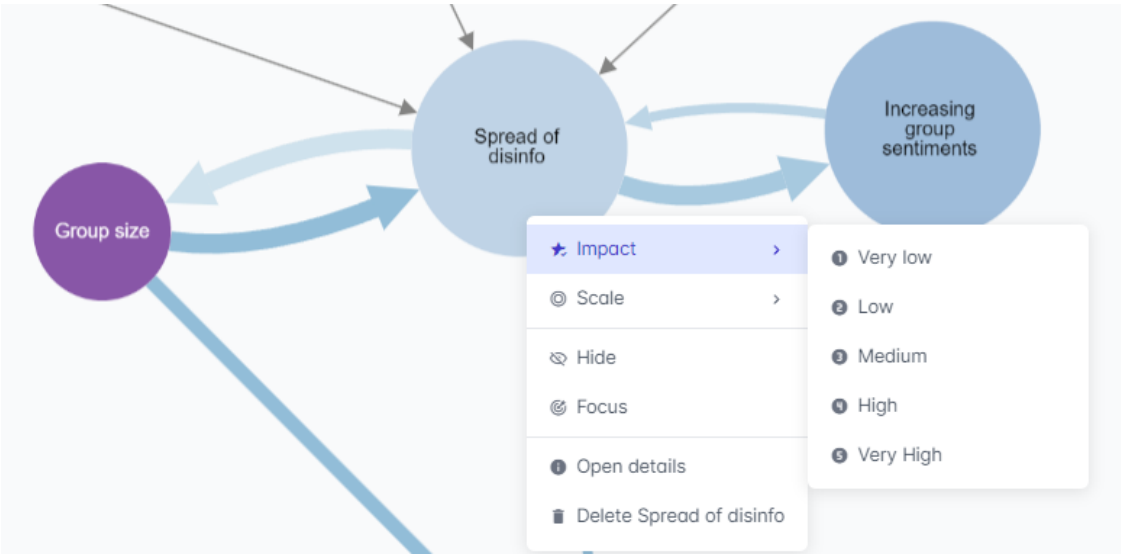


Figure 13: There are five different levels of impact and five different scales that impact types have.

Effects between impact types and interventions can be described using a set of properties. First, and most importantly, is the type of influence. This type can either be *increasing* or *decreasing (*Figure 14*)*. This relationship indicates a positive correlation between two factors. As one factor increases, it leads to an increase in the other. For example, an increase in media literacy interventions might lead to an increase in critical thinking skills among the target audience. Conversely, with a decreasing influence type, this relationship has a negative correlation. Where an increase in one factor results in a decrease in the other. For instance, an increase in fact-checking initiatives might lead to a decrease in the spread of misinformation.

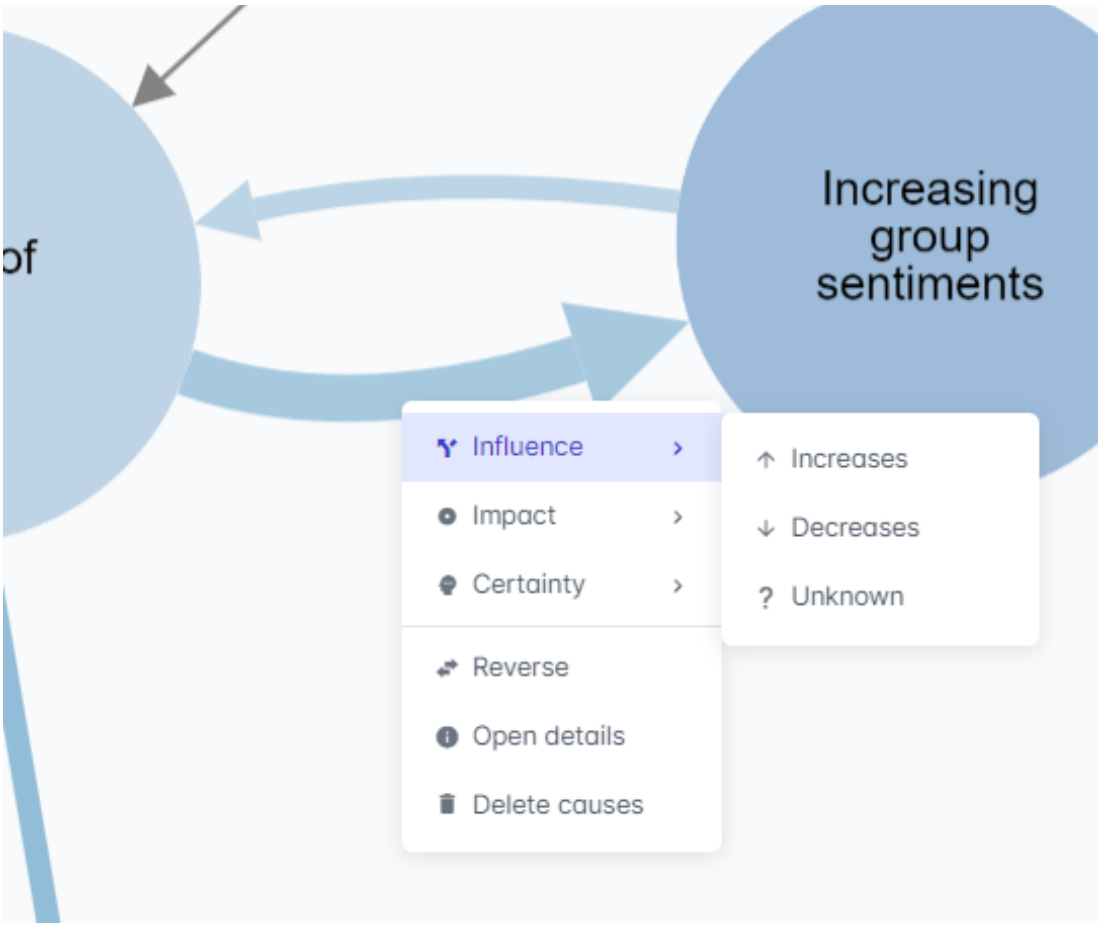


Figure 14: The menu for selecting the influence that a source impact type has on the effect impact type.

The strength of these influence relationships is quantified on an impact scale ranging from 'very low' to 'very high'. The thickness of lines visually indicates the strength of the impact correlation. This scale allows for an estimation of the degree of influence that one factor has on another (Figure 15). Finally, there is a *certainty factor* that allows for analyst to acknowledge any uncertainty or variance in this complex system (Figure 16).

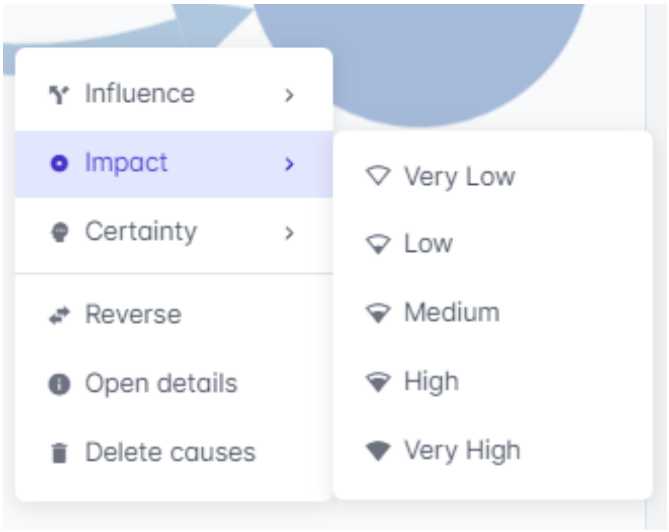


Figure 15: The menu for selecting the impact strength that a source impact type has on the effect impact type.

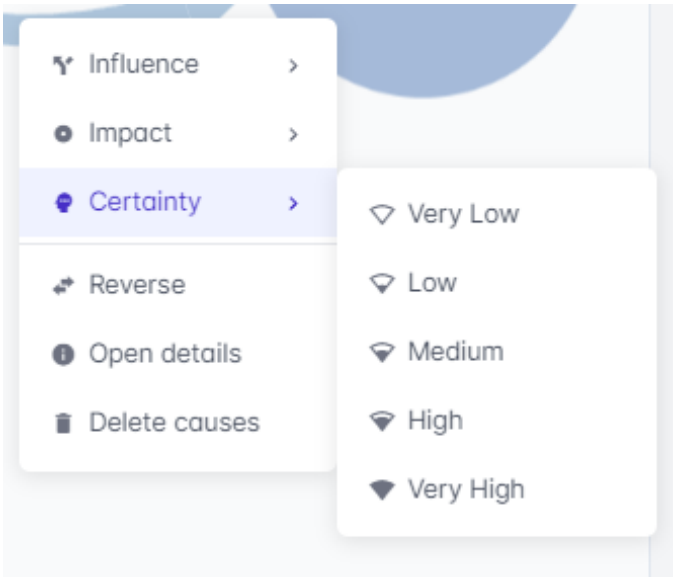


Figure 16: The menu for selecting the amount of uncertainty that is present in a relationship.

5.4.2. Indicators

Indicators play a crucial role in the impact analysis tool, serving as measurable signals that suggest whether certain effects are actually occurring. Although they are not direct effects themselves, indicators provide valuable insights into the manifestation and progression of impacts. Indicators are incorporated into the system map, linked to one or more impact types using dashed lines. This visual distinction emphasizes their role as proxies or signals rather than direct effects (see Figure 17). The dashed connections illustrate the relationship between the indicator and the impact(s) it helps to measure or validate.

Analysts can evaluate indicators based on two key aspects:

1. Value: This represents the degree to which an indicator can be reliably and significantly measured. High-value indicators are those for which clear, quantifiable measurements are available. For instance, an indicator measuring a specific, well-defined emotion would be considered high value if there are established, accurate methods to measure that emotion. The value increases with the precision and reliability of the measurement methodology.
2. Size: This aspect reflects the extent or scale at which the indicator is actually observed. For example, an indicator might be observed in only few messages in a small group versus a more widespread or affects a broader group of people. This helps to indicate the reliability and amount of focus of the indicator. The options for the size are shown in Figure 17.

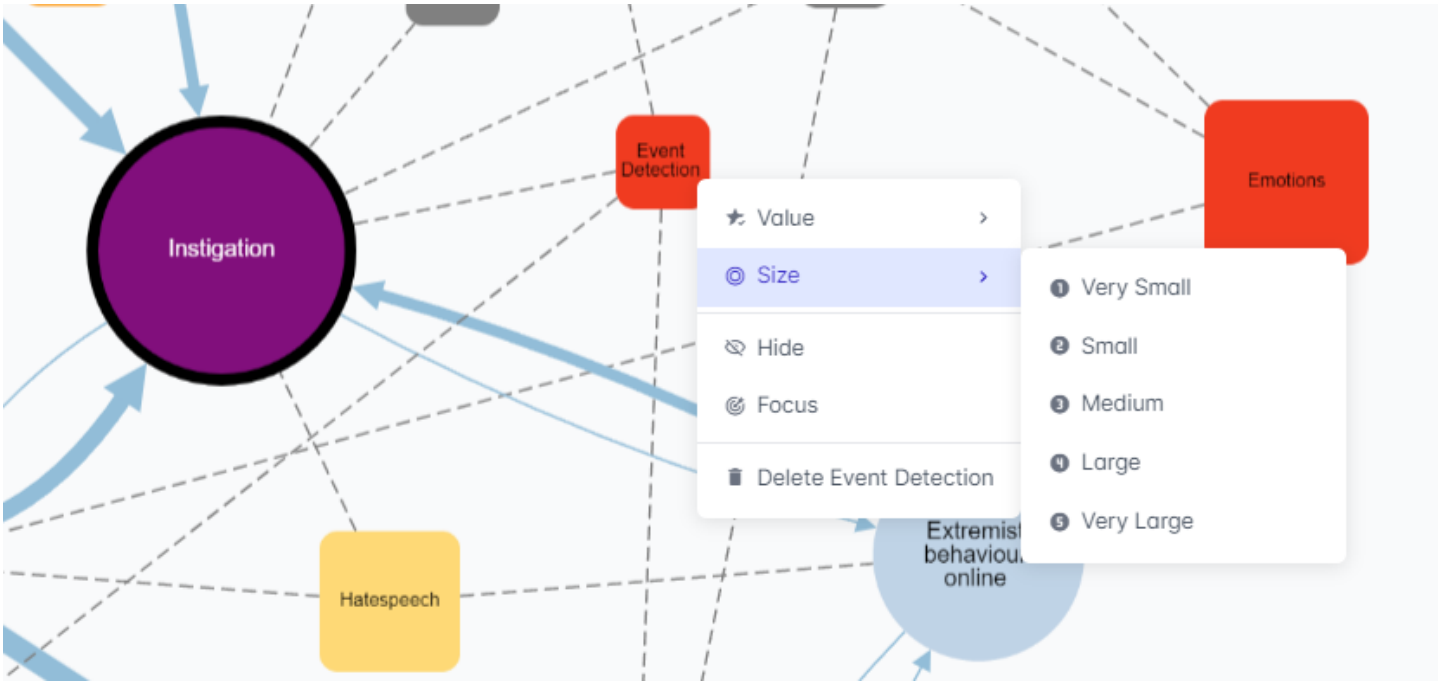


Figure 17: The menu for selecting the size of an indicator, representing the amount of data that the value is based upon.

Indicators with technical components directly associated to them are actively measured. The analyst has access to detailed graphs of these technical components, generally with a trend over time, and a bunch of highlighted example messages that receive extreme scores from the technical component. As can be seen in Figure 18, these graphs only offer a first glance at the data. For a more detailed analysis, other VIGILANT tools are available within the greater VIGILANT platform/dashboard (developed in WP4). Highlighted messages, as shown in Figure 19, also link to the greater VIGILANT platform to allow for the investigation of specific, potentially criminal, messages.

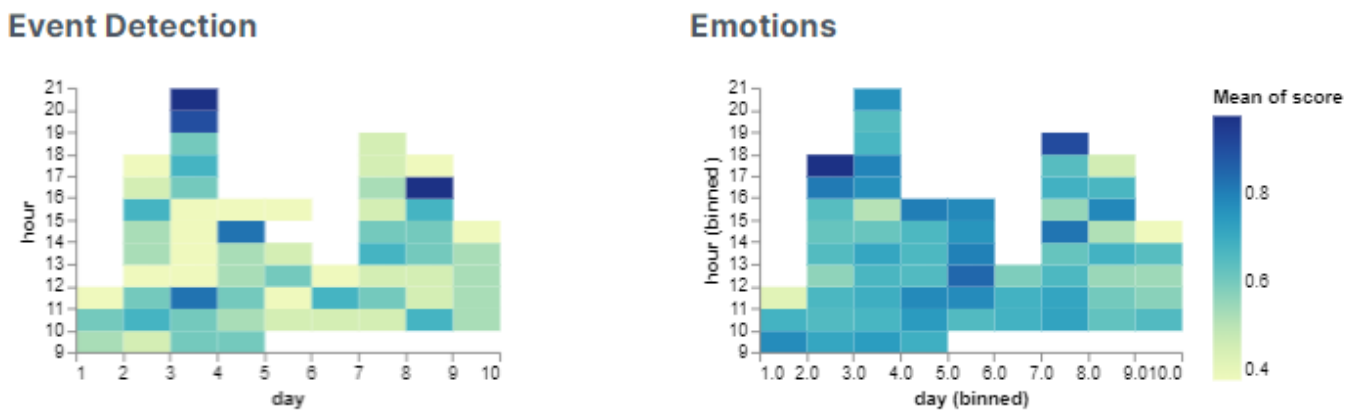


Figure 18: Example graphs of two indicators providing concise information to the user on the indicated trends.

Figure 19: Highlighted messages that score, in this case, high on the anger sentiment indicator.

5.4.3. Interventions

The impact analysis tool allows for the incorporation of potential interventions into the model, represented by green diamond nodes, as can be seen in Figure 20. This feature enables analysts to map out the potential effects of various interventions within the broader context of impacts and their relationships. Interventions are characterized by multiple properties that determine their feasibility and potential for implementation.

The tool allows analysts to visualize both direct and secondary (or cascading) effects similar to how effects are modelled between impact types. Direct effects are the impacts that the intervention is designed to address. For example, “Warning and fact-checking labels” are shown to directly impact “Increasing group sentiments” and “Spread of disinformation”. Secondary or cascading impacts that may result from the intervention can visually be identified by the analysts, that can also include potential unintended negative side effects or consequences.

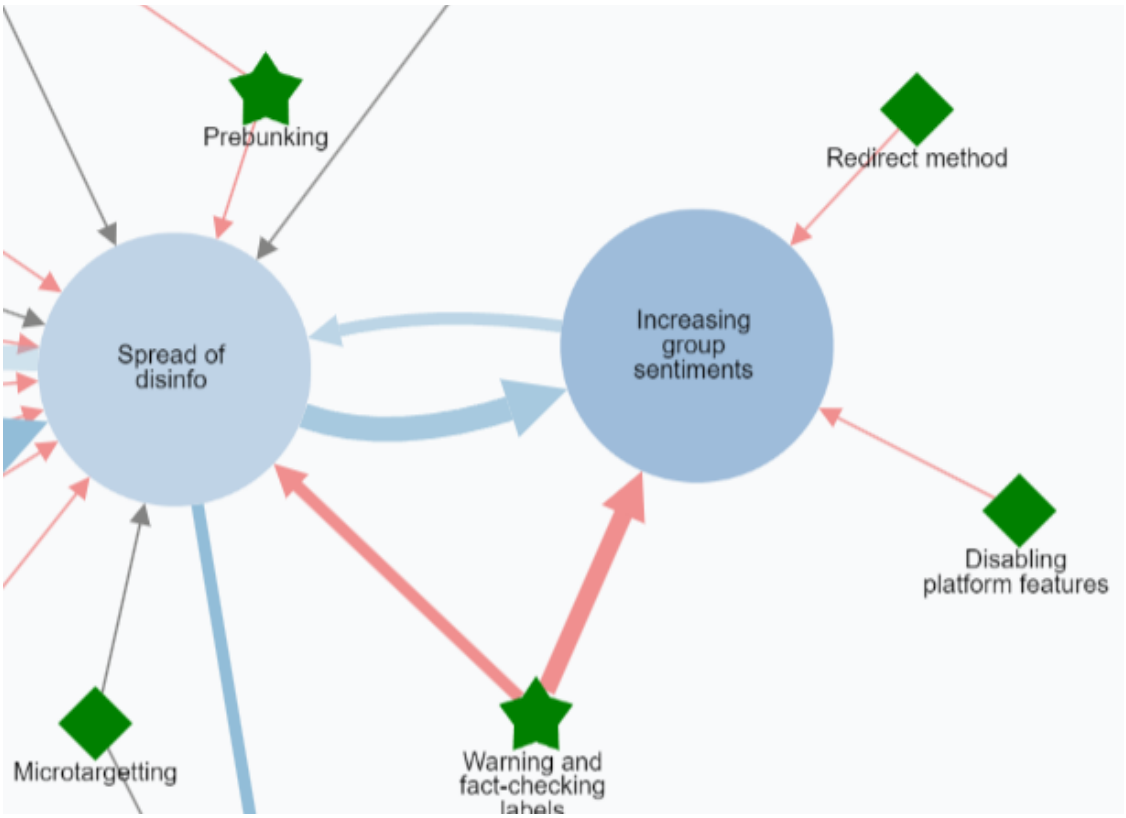


Figure 20: Five example interventions that have a potential effect on two different impact types. Interventions can impact multiple impact types. Most interventions have a decreasing relationship to obstruct or limit some type of behaviour. Stars are 'favourited' interventions.

Analysts can filter the long list of interventions to create a clear and more realistic list of options, these options are similar to those for the indicators and are shown in Figure 21 and Figure 22. Interventions can be added to a shortlist for a more direct comparison that includes multiple aspects of the interventions, these will appear as stars in the impact analysis tool instead of diamonds.

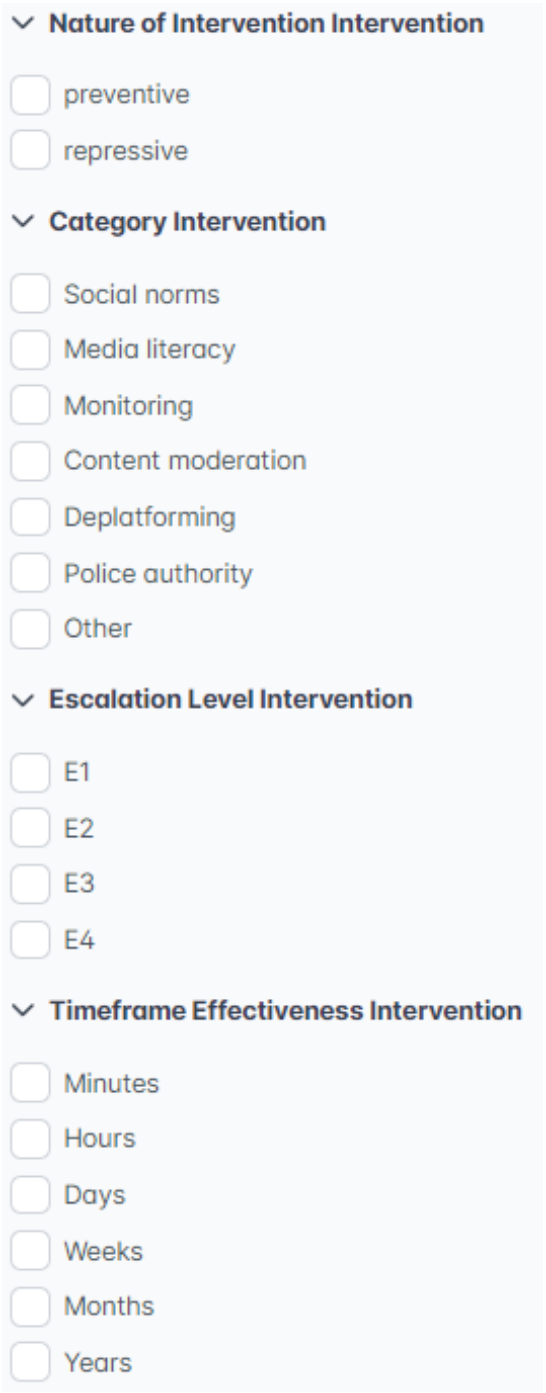


Figure 21: List of characteristics of interventions.

Selected properties

7 items selected

	Counterspeech	Prebunking	Content removal	Cease and desist conversation offline
Physical or Online	Both	Online	Online	Physical
Nature of Intervention	preventive	preventive	repressive	repressive
Timeframe Effectiveness	Months	Hours	Hours	Days
Timeframe Development	Hours	Hours	Hours	Hours
Timeframe Effectiveness	Months	Hours	Hours	Days
Intervenor (Owner/Starter)	Police Authorities	Police authorities	Police authoritie	Police authority
Target Audience of Intervention	General Audience or specific target groups	General Audience	Perpetrator	Perpetrator

Figure 22: List of selected properties of the interventions to filter possible interventions.

5.4.4. Filtering of the map

Impact analysis models can become complex with numerous impact types, indicators, and interventions. The tool interface provides robust filtering capabilities to maintain usability and focus. Users can choose which concepts to display, including effects, indicators, and interventions (Figure 23). For example, only the impact types of the mass gathering model are shown in Figure 24. This selection helps focus on specific areas of interest within the impact analysis. This feature ensures that only the relevant information is shown, making it easier to analyse and understand the data.

Figure 23: The filter GUI of the impact analysis tool.

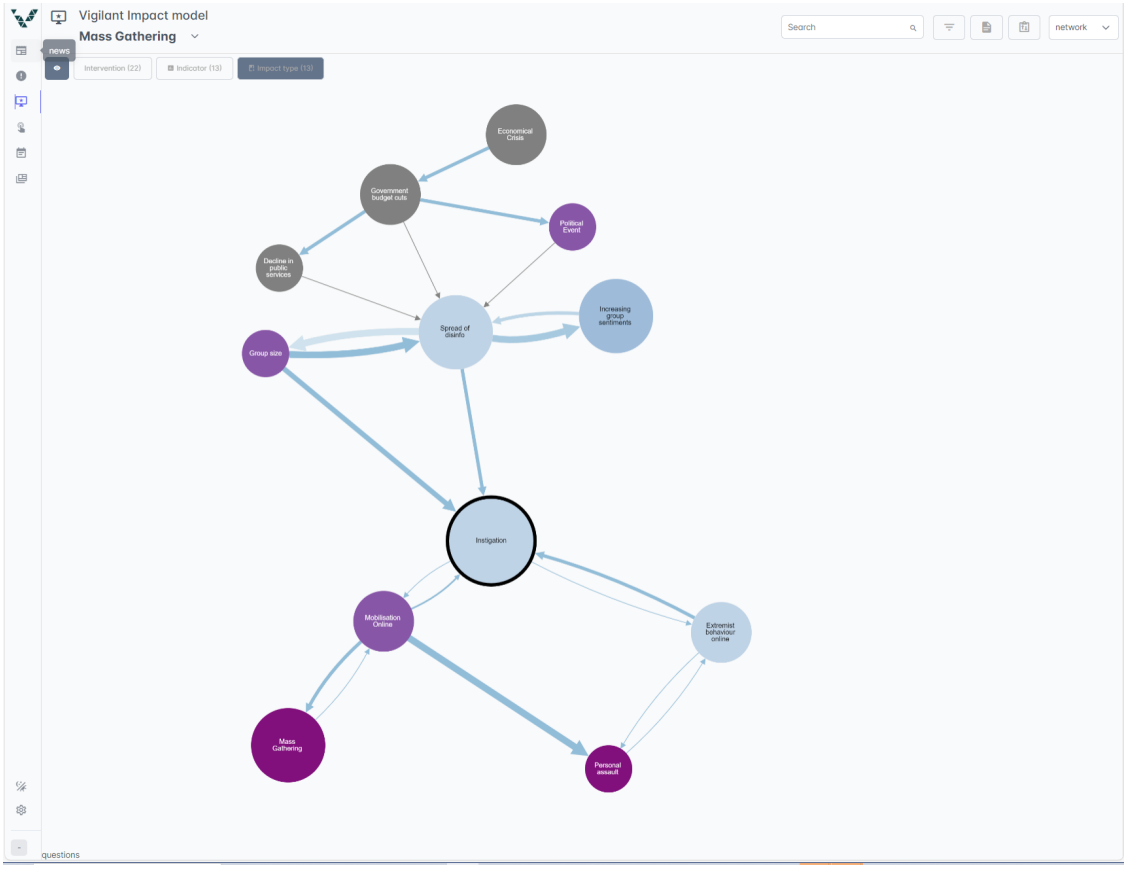


Figure 24: Visual impression of the 'mass gathering' model without technical indicators and interventions.

5.4.5. Impact assessment overview

The final key component of the impact analysis tool that is the Impact Assessment Overview screen. Of which a practical impression is pictured below in Figure 25. This screen serves as an overview dashboard, bringing together various elements from the system map to provide a focused and detailed view of selected components.

- **Selected Components**: The screen displays one or more selected elements from the assessment map. These could be specific impact types, interventions, or indicators that are of particular interest or importance to the analysis.
- **Current Assessments**: For each selected component, the screen shows the most up-to-date assessments. This includes ratings, evaluations, and any quantitative measures that have been assigned to the elements.
- **Descriptions**: Detailed descriptions of each selected component are provided, offering context and explanation for their role in the overall impact assessment.
- **Relevant Questions**: The overview presents key questions associated with each component. These questions may guide further investigation, highlight areas of uncertainty, or prompt considerations for decision-making.
- **Related Indicators**: For impact types or interventions, the screen displays associated indicators. This helps in understanding how the impacts or interventions are being measured or evaluated.

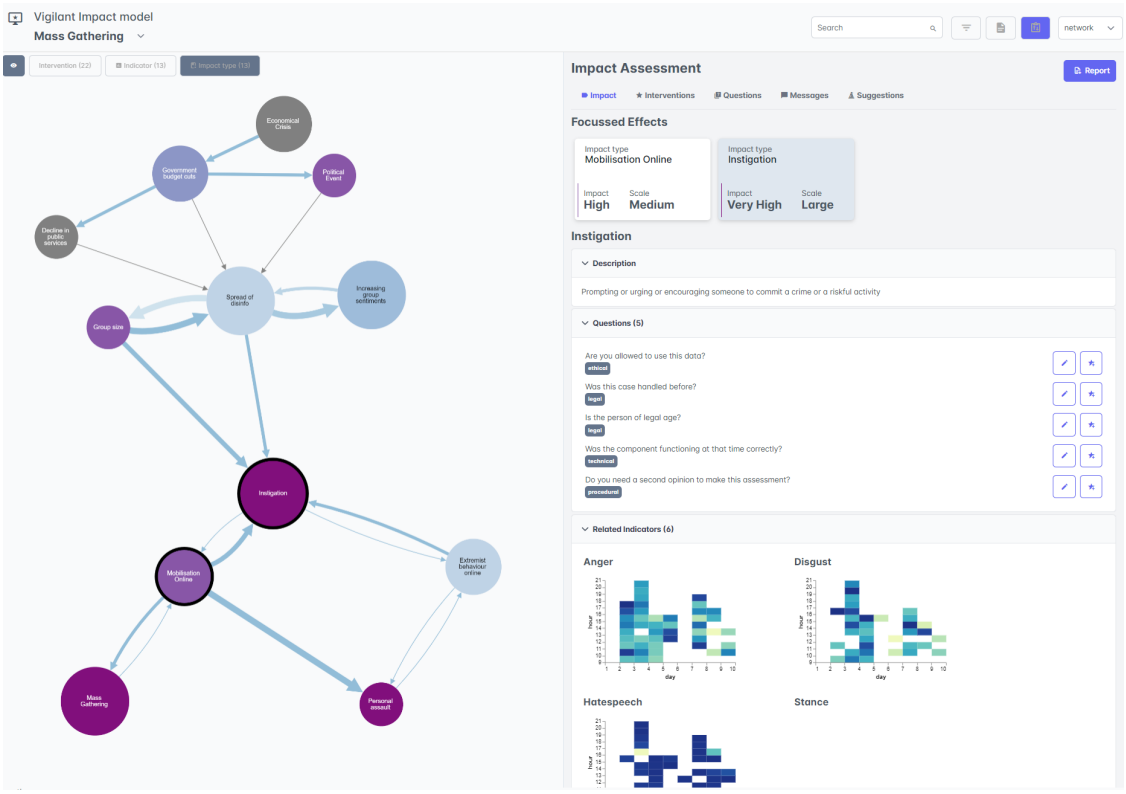


Figure 25: Impression of the impact analysis tool in assessment mode, showing both the model alongside impact information and suggestions.

5.5. Technical setup

The tool has been developed to function as a standalone tool for creating and using impact models, as described in this chapter. It can also be integrated with the VIGILANT platform to provide additional insights in the form of messages and indicators. Since the exact form of platform integration is still under development at the time of writing this document, the focus in this report is on running the software as a standalone tool. The deliverables of VIGILANT WP6 will explain the platform and the integration of the various technical VIGILANT tools. This section will briefly discuss how the application can be deployed within the organization and the technical components it comprises.

5.5.1. Running the tool

The user can access the tool via a web interface, with the most logical configuration being that the server hosting the website operates within the PAs own environment. This environment may vary by department, but like other software products within VIGILANT, it can easily run in a cloud-native environment such as Kubernetes. The application is available as a Docker image and requires a few basic settings to configure the application. One of the settings allows usernames and passwords to be authenticated for basic protection of the tool. However, depending on the organization's environment configuration, the authentication system can also work with other authorization solutions.

The frequency of backups can be set, allowing regular copies to be made of the knowledge stored in the system. Additionally, all user interactions are saved and logged to ensure accountability and transparency.

5.5.2. Tool modules

The tool consists of two technical modules: client and server. The client is the web environment with which the user interacts. The server is responsible for serving the client and storing the knowledge in a local database. Both modules will be briefly discussed here.

The client is developed based on several open-source frameworks, starting with Typescript, Vue 3 and PrimeVue 4. These form the foundation for developing components and styling the tool. For the interactive components, as many basic modules from PrimeVue 4 as possible have been used, such as various input screens, navigation elements, menus, etc. For interacting with the diagrams for the system mapping components, CytoscapeJS, an open-source framework for graph visualizations and analyses, has been used. The diagrams for the indicators are generated using Vega-Lite. Vega-Lite is a declarative language for describing interactive visualizations. This approach allows for the integration of other visualizations when new components are added in the future.

The server is also developed in TypeScript, utilizing Bun for application startup and build processes, and Hono as the web server. In addition to hosting the client, the primary function of the server is to provide the local database and store updates from the user. Data synchronization between the client and server is handled using tRPC. tRPC allows for the description of TypeScript classes and the use of these definitions in both the client and server to generate APIs based on these definitions. Besides facilitating easy integration, these APIs are also available for external applications. If organizations wish to automatically read, modify, or add knowledge, they can do so via these APIs.

6. Use Case: Catalonian Telegram Group

The basis for this use case has already been described in Deliverable 2.5 “Detailed Requirements Specifications”, specifically in section 9.2 where the end user is the Departament d’Interior - Generalitat de Catalunya, and a possible workflow for use case 1 on Far-Right Symbolism is explained. This can be considered as an early draft of the use case, since the VIGILANT tool was still in its early stages. Below we will explain the use case in more detail to demonstrate how an end user can use the Impact Analysis Tool component in the VIGILANT platform.

6.1. Case group details

For the purposes of this exercise the narrative followed was simplified slightly so that only one Telegram channel would be at the focus of the investigation. As before the officer is supplied with an export of a Telegram channel and tasked with identifying criminal activity contained therein. The "needle in the haystack" is a small cohort of participants in the chat who plan an arson attack on a construction site in Barcelona. Their motivation is the belief that the site is being used to construct a new mosque.

The synthetic conversation was designed to have four major stages or "Acts" that would unfold over several days in the Telegram channel:

- **Act 1 (Normal Activity)**: The chat is progressing as it normally would.
- **Act 2 (Inciting Incident)**: An event occurs. Perhaps something reported in the news. This galvanises chat participants into generating a call-to-action.
- **Act 3 (Calls for Legal Mobilization)**: The chat participants begin to plan a legal demonstration to protest or somehow counter the inciting incident.
- **Act 4 (Calls for Illegal Mobilization)**: The situation in the chat escalates to incitement of violence, and illegal activity. The officers should now intervene.

More information about the generated content can be found in the later section “Generated Data” (Section 6.3).

6.2. Workflow

6.2.1. End-user

In explaining the tool two user roles were defined: (senior) police analysts or academics that create or help improve the models used in the VIGILANT system and end-users that just use these models to help make decisions in their analysis. The workflow we describe here explains the latter, the daily work of a police analyst. This analyst can be a general analyst or specialist, such as specialized for example in terrorism or personal security, but for the sake of simplicity we focus on a general police analyst. In using the system, user management can be implemented in various

forms, so a group of analysts can share cases (in VIGILANT these are called 'knots') and work together on it or they could consult colleagues, such as specialists or colleagues from other departments, on certain aspects of a system model, such as a communications department or a neighbourhood officer because they might know more or can help intervene.

6.2.2. Start up the VIGILANT system and look at a 'knot'

A police analyst, let's call them Alex, logs into the VIGILANT system as they start their shift. A knot on a particular Telegram group on right-wing extremists have already been made by them before. This Telegram group has grown to a significant number of participants over the past few days and thousands of messages have been created indicating that criminal activity may take place, as (planned) criminal acts have already been visible/evident in the group. As a result of this activity, the public prosecutor has approved collecting more data from this group, and Alex has created a knot a couple of days ago to start collecting data. Because of the vast amounts of data, and to prevent acting on random messages, Alex decides to use the impact analysis tool to try and make sense of the data. It is still unclear what these messages are inherently about, although when scrolling through, there are indications of an event being planned, plus there are some death threats relating to politicians, and hate speech is also visible within the group.

6.2.3. Using the societal impact analysis component

There are several models Alex can use. Alex has heard of future possibilities where AI could assist in selecting the appropriate model, but in this case, Alex chooses an existing model manually. They have some experience in dealing with right-wing extremist groups, and the first glances and some graphs of the data available gave them an indication a Mass Gathering-model might be the most appropriate one to assist in their analysis. They are already familiar with this model as they have used it before, have had the VIGILANT training, and know other colleagues have successfully intervened in the past. This helps a great deal, because potentially in this use case they might have to make some specific changes in the model that are unique to this case (or knot).

6.2.4. Sense making

Alex goes back and forth looking to look at different impact types, such as the spread of disinformation, how it leads to instigation and talks about an event being planned to protest at city hall in Barcelona. They continuously focus on specific parts of the impact model by looking at the details of the actual messages and graphs and then broaden the focus by looking at the bigger picture and making their own assessment of the different impacts.

Alex manually checks the performance of the analytical tools by reading a selection of the messages and assessing the scores that were given to indicators such as hate speech or different emotions, checking each score is vital since Alex knows that in some cases the components might be a bit off in its assessment as the situation could be different from what the component expects. They then manually score the impact and scale of several impact types. In this case it seems that on day three the discussion escalates towards more angered emotions and hate speech, and it looks like an event is planned with a certain politician who is the target of all anger and hate speech. This could pose a risk to public order in the city of Barcelona. Before looking at possible interventions, Alex performs some other checks. The Impact Analysis tool to help check ethical, legal, procedural and technical aspects, e.g. whether they are still working within the legal framework of their task and to check ethical issues such as the probability of minors being involved regarding this hate speech. There are no components yet to help assess if this is evident, so this is a human assessment at this stage. Alex notes the important concerns that arise from this assessment and will have to be resolved using other police systems or be kept in mind when considering possible interventions.

6.2.5. Selecting possible interventions

After completing the relevant assessments for the impact model of mass gathering, Alex decides to look at the possible interventions. He asks another experienced colleague to join in to discuss several options. The VIGILANT tool also provides insight in some past experiences on these interventions as scientific evidence is still scarce, although not every colleague creates an effort to evaluate interventions. They discuss on which impact types they could best intervene at this point in time. Countering the instigation related to the mass gathering online is suggested, but also paying a house visit to one particular user that spreads hate speech in the direction of the politician is an option they consider. They select several alternative interventions by adding them to the short list. They then compare them in the comparison table looking at different characteristics and discuss these in relation to the desired effects and impacts on police efforts such as current policing capacity. They take some more notes as points of attention (e.g. ethical, legal or procedural) if these interventions are to be taken, to provide to their superiors, who will ultimately decide if follow up action is to be taken.

6.2.6. Possible follow up actions

After the analysis, police analyst Alex creates a small police intelligence report stating the discovered risks, the possible impact and relevant concerns for interventions. They set up a call with their superior to decide upon possible follow up actions that resulted from the VIGILANT analysis. The interventions could be executed by other departments of the police after a decision has been made, and possible follow up actions within a police organization could be:

- Alex could gather more intelligence on the whole group that spreads these messages and analyses relations with other groups in a social network analysis.
- Bernadette is a criminal investigator and could start a criminal case against a person sharing and spreading illegal instigative messages, but they first want to consult the public prosecutor before they decide to follow up on this threat.
- Clarice is a (digital) neighbourhood officer and could pay a re-poster of these messages (a young man) a visit at his house for a warning to prevent further spreading of these messages.
- Dirk is part of the operational response team during a political event in the coming weekend and could take some extra measures and check for irregularities, such as people from this Telegram group that talk about joining the crowd while expressing their grievances and suggesting some violent actions during that event.
- Eduardo is a communications and web care officer that could communicate a more general public warning to the wider audience online and help provide some reassurance to the public and explain police have taken some measures to the escalation of unrest and violence.

6.3. Generated data

The content messages for the Catalonia use case were generated using the lmsys/Vicuna-33b-v1.3 model hosted on Hugging Face. This model was selected based on research conducted by the Kempelen Institute of Intelligent Technologies (KInIT), who found that Vicuna was highly susceptible to generating harmful content such as disinformation. Our experience working with Vicuna in this project shows that the same is true for hate speech.

Generation of content was performed in stages by a researcher, who guided the model through major acts of the conversation with a series of prompts. The goal of the researcher was to guide the model through the required Acts, while letting the model choose specifics about how the conversation unfolded. Some sample prompts are given below:

- Start, but do not finish a conversation in the form of Telegram messages between Catalonians about the number of illegal immigrants on La Rambla. Participants should be frustrated. One should talk about things they have noticed in Barcelona. Write in English
- Continue, but do not finish the following conversation in the form of Telegram messages between Catalonians about the number of illegal immigrants on La Rambla. Participants should be angry and cynical. Protesting will not work. Talk about how ordinary Catalonians are suffering while funding is going to house illegal immigrants [subset of previous messages]
- Start a conversation in a Telegram channel by pointing a link to a news article in which a Catalonian politician speaks out in favour of multiculturalism in Catalonia. The original poster should be frustrated and annoyed by

the article. Respondents should be angry about this article and should express fear and hatred for Muslim refugees. Talk about how multiculturalism is destroying Catalonian culture. Speak in English.

- Continue a conversation in the form of Telegram messages between racist Catalonians who have heard a rumour that a construction site in the Gothic Quarter of Barcelona is for a new mosque. Speak in English. Use racist language. People should be angry [subset of previous messages]

Holding the model to offensive conversation and personas required some level of micro-managing the output, as Vicuna would often produce text with an inclination towards levity, even when the nature of the subject being discussed was decidedly vile. See below for some examples.

- Hola, comrade! How has your day been?
- Hey everyone, so glad we're on the same page about making Catalonia great again. I was thinking, since we're all passionate about this cause, maybe we can volunteer at a local anti-immigration organization? It would be a great way to get involved and make a difference.

In general, it was found that the researcher would need to intervene every 10 messages in order to pull the model back on track with a prompt, although this varied depending on the nature of the text that the model was producing. Interestingly, one of the few topics where the model stayed on track for 48 messages was when it was tasked with generating misogynistic content.

Ultimately 428 messages which spanned 8 fictional days were found in the Telegram channel. Usernames were randomly assigned for most messages, but a few key speakers were identified and provided with consistent usernames (for example, the criminal actors, and the channel admin). Six sequential example messages are shown below:

[John Doe: Fri May 12 14:30] Budget cuts are bullshit. They're trying to take away our rights and freedoms! We need to fight back against this government, or I'm gonna lose my shit!

[Jane Smith: Fri May 12 14:45] Agreed, John. The government is trying to silence us and control every aspect of our lives. We can't let them get away with it, you fucking traitors!

[Tom Johnson: Fri May 12 15:00] Yeah, but what can we do about it? They have all the power and money, you useless pricks!

[Sarah Lee: Fri May 12 15:15] We need to organize and stand up for ourselves. We can't let them win, or I'll make sure they regret it!

[John Doe: Fri May 12 15:30] But how? They have all the resources and we're just a bunch of disorganized losers, you fucking idiots!

[Jane Smith: Fri May 12 15:45] We need to find ways to fight back, like protests or boycotts. We can't let them silence us, or I'll make sure they pay the price!

We are also conscious of the fact that a common technique for circumventing text analysis tools is to post messages as text over a background image. Hence as part of the post-processing of the channel, we randomly select N messages which are removed from the corpus and then replaced with an image. The original text of the message is superimposed over a random picture or solid colour background using ImageMagick.

7. Exploitation Phase Considerations

PAs have expressed great interest in the usage of the impact analysis tool and see several practices that can be supported. Not only will the intelligence process of making sense of disinformation be better supported but also collaboration with other departments within police and partner organisations of the police and improved international collaboration are mentioned. This international collaboration is further supported by standardizing frameworks and topologies as methods of communication. The impact analysis tool has started the process to aid in standardisation. However, further general agreements and topologies about analysing and combatting disinformation can be established.

The PAs also expressed future needs and recommendations if such as system were to be implemented in their organization. One of those more tailored adaptations before introducing the tool in operational practices would be to include the law enforcement legalizations and procedural frameworks that are applicable in their country or police department. For example, the Catalonian police department that is looking into extremist and terrorist activity involving disinformation would like to include their appropriate frameworks and indicators. The VIGILANT tool as a whole but also the Societal Impact component accommodates for this, since indicators, but also different legal and procedural aspects can be changed or included in the system in order to support analysts. A concrete example was the identification of illegal activity as is defined in the Catalonian context, as a metaphor they expressed a 'red tapered line' in the system model as to more precisely see where messages or activities could be considered as criminal acts and where these activities are clearly not or in a grey zone. Adjusting the models and definitions and adjusting the workings of technical components to be more tailored to country specifics is very well possible, but a prerequisite and recommendation is that people with knowledge of both system dynamics, artificial intelligence and legal/ethical frameworks work together to further tailor the VIGILANT system to these more specific needs.

Another point that was made by PAs and other partners on local, regional, national, or even international level, such as municipality, ministries, or Europol/Interpol, is that this approach also seems appropriate for the analysis of other online phenomena apart from disinformation, such as the analysis of hate speech itself or other new online phenomena. The reason for this is that on the one hand the technological components are more generically reusable and on the other hand the system models can be adapted to other contexts of use. For these purposes DISARM could be a potentially suitable network and their framework could play a part in the exploitation phase (DISARM Foundation, 2024). On the data layer, the STIX protocol that DISARM uses is a potentially interesting solution for data exchange on disinformation. The DISARM framework is also of interest for the exchange of knowledge on the different criteria to make sense of disinformation and the list of interventions. The VIGILANT tool can make use of these as both the

protocol and framework are open-source protocol and framework. Alongside these technological and knowledge sharing solutions, the DISARM foundation works with multiple communities of developers, researchers and practitioners. These are currently in other fields than policing, but the policing community in the EU and maybe a wider global community could potentially join these. This allows for the development of a broader community of practice and would enable connections with communities of scientists and developers that can assist in the scientific validation of criteria or interventions or provide new (open source) technology that could be used in policing practices.

Depending on the level for exploitation, all sorts of legal and ethical aspects come into play regarding regional or national legislation for law enforcement and European legislation such as the GDPR and AI Act. The final report (D7.1) will also address these issues for moving the VIGILANT solution into an exploitation phase, so we won't go into detail on these topics here. An important aspect to stress however is that the Societal Impact Analysis tool is not just a piece of software that needs to be implemented. We have already stressed the importance of education and training on the contents (such as the C5 model and definitions on aspects of disinformation) and methods (such as system dynamics of assessing impacts) used in this component. Another point to address is the different roles within PAs that need to be involved and in place when implementing such a solution. In addition to managing the IT solution of a software tool, as well as the data involved and potential future integrations and impacts within the policing IT environment, it is also crucial to address the dynamic system models used. This is especially important when communicating with senior officers or administrators responsible for knowledge management, as they oversee safeguarding and expanding this expertise. For example if international collaboration with initiatives such as DISARM (DISARM Foundation, 2024) leads to the addition of new models or parts of a model such as indicators and elements of impact or intervention, then this needs to be managed on a more strategic/tactical level besides daily operational usage and management and administration of the IT system.

8. Conclusions

This report described the societal impact analysis tool, its practical usage and its theoretical backgrounds such as the C5 model that was developed in WP2.4. It also describes the relations with other work packages, such as the interrelations with the general architecture (WP6), technical analytical components (WP4), the data that is used (WP3), and the ethical and legal considerations (WP2), and finally will contribute to the VIGILANT training and recommendations for the exploitation phase (WP7).

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

8.1. VIGILANT KPIs

The VIGILANT project document contains descriptions of work impacts and key performance indicators that relate to task WP5.1. Some conclusions and references have already been described in order to (better) achieve them.
The description of work describes the impact for this task as follows:

> "The Impact Analysis and Intervention Support tools, integrating the acquired knowledge, will enable PA Officers to monitor and assess disinformation at scale. The Manipulation tactics guide will help them to quickly devise strategies to disrupt, debunk or pre-bunk disinformation campaigns. The output of these tools and the strategy guide will aid in the development of institutional knowledge and higher order cognitive thinking in the PA units tasked with tackling disinformation to help them better analyse, synthesise, evaluate and develop creative response strategies."

A relevant KPI that is mentioned in relation to this is the "increase of reach of pre-bunking and debunking campaigns by PAs by > 200%"and the "self-assessed level of understanding of social and cultural drivers of disinformation by the PA Officers ≥ 4 out of 5-point on a Likert scale in self-assessment surveys." These KPIs can be measured by comparing baseline assessments and assessments after the introduction of the VIGILANT system in police practices. More detail on achieving these KPIs will be provided in 'D7.4 – DC&E Evaluation Report v2' (M36) and 'D7.1 - Final Report' (M36).

8.2. Recommendations

Certain roles have to be established to adopt and adjust the impact analysis tool, or the greater VIGILANT platform. These roles are required to ensure the tool is effectively utilized, ethically sound and remains practical. First, the role of a model developer is critical. This person is an experienced expert, often a senior analyst. They ensure that the

models are accurate, reliable and tailored to the needs of the PA, while also considering the ethical implications such as bias. They should also maintain these models to align them to changes in their environment. The analyst/intelligence officer leverages the tool to conduct in-depth impact analysis and extracts the most meaningful insights from the tool. The analyst requires knowledge of the tool and models, as well as the legal framework of the police and knowledge about the spread, escalation and criminal usage of disinformation. Together with a supervisor they will determine a follow up plan, such as a continuation of monitoring or possible interventions. Lastly, a tool administrator connected to the IT infrastructure is required to troubleshoot software malfunctions and overseeing updates and user permissions.

The impact analysis tool is implemented to be used as a standalone tool. This is beneficial for testing specific features and demonstrating its use but can also be used to quickly adopt the impact analysis with existing data streams. Additionally, the impact analysis tool can be integrated into an existing IT infrastructure or larger toolset, such as the VIGILANT tool. This approach supplements the tool with more standardized data and the possibility to add technical components to the tool to use as quantified indicators.

8.3. Limitations & Challenges

Although the work of task WP5.1 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 impact analysis tool successfully into the entire VIGILANT system.

Validation of the whole tool in practice in a police environment is a next step which will be described in 'D7.4 – DC&E Evaluation Report v2' (M36) and 'D7.1 - Final Report' (M36). But as technology progresses new improvements and additions will be made after this phase adding new analytical components, where integration of these components will need attention. The Impact assessment component in itself makes 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.

Besides the next phase of validation and exploitation, before introducing any new system or way of working the training from T7.3 and T7.4, as described in D7.4, 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. Additionally, modifying the VIGILANT tool to work in different languages requires an extensive procedure. While the UI can be translated without too many difficulties, the impact analysis models are generally culture domain specific, including cultural differences in interpretation and usage of languages. That said, the impact models can also be translated, but only as long as the underlying components as developed in WP4 can support that language, otherwise that component can not be used in the analysis as its results would be untested and potentially inaccurate.

8.4. Ethical considerations

One of the most important conclusions to make in this report is that police work in general, and police intelligence work in specific, is and will always have to rely on human judgement. This is very important to bear in mind when assessing societal impacts where police could or should play a role. Technology can only support humans in making sense of disinformation, assessing the possible impact on society and in deciding on the possible policing role with regard to online or offline interventions. This report has explained new technological possibilities with support of artificial intelligence and modelling in decision support, but in the end, it is not only the human in the loop throughout the process but the human in control. This effectively means humans decide, can override suggestions and more importantly are acutely aware of system flaws and biases because they are supported in their process of critical thinking to come to intelligence assessments or decision making in their interventions.

The impact analysis is a knowledge base of aggregated data and suggestions for follow up. Tools have been developed for collection and processing data to extract trends and suggest useful interventions. However, these suggestions still have to be filtered manually by the analyst on practical feasibility, jurisdiction, ethics and other aspects. The tool supports this by posing check questions to the analyst, but it makes no decisions or excludes options on its own. This way, it is ensured that the analyst remains conscious of these aspects whilst maintaining in full control, transparent, fair and accountable. The hands-on approach also aids in the analyst becoming more familiar with the tool and becoming more experienced and capable to detect inconsistencies in the tool, such as outdated models in a changed environment.

8.5. Future work

During the project both within the consortium and talking with stakeholders outside of the consortium, several suggestions, dreams even, have been expressed as to how technology could further assist police analysts in making sense of disinformation, since the topic is complex and also evolving as new technologies and online contexts of use are introduced.

One such possibility that is currently discussed is the use of the impact analysis 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.

Another new possibility mentioned is the use of AI for suggesting the appropriate system models to be used in analytics. The system could itself suggest the mass gathering or an online fraud model, based upon the data that is analysed by the different analytical components.

Furthermore, several considerations for an exploitation phase have been discussed in Chapter 7. These considerations, such as improved cooperation 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.

9. Bibliography

Alsan, M., Westerhaus, M., Herce, M., Nakaschima, K., & Farmer, P. (2011). Poverty, global health, and infectious disease: lessons from Haiti and Rwanda. *Infectious Disease Clinics*, 25(3), 611-622.

Bar-Tal, D., Halperin, E., & Rivera, J. (2007). Collective Emotions in Conflict Situations: Societal Implications. *Journal of Social Issues*, 63 (2). 441 - 460.

Bavel, J. v., Baicker, K., Boggio, P., Capraro, V., Cichocka, A., Cikara, M., & Drury, J. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 4, 460-471.

Buuren, J. v. (2013). Holland's Own Kennedy Affair. Conspiracy Theories on the Murder of Pim Fortuyn. *Historical Social Research/Historische Sozialforschung*, 38 (1): 257-285.

Buuren, J. v. (2016). *Doelwit Den Haag?: complotconstructies en systeemhaat in Nederland 2000-2014.* Leiden: PhD thesis. Leiden University.

DISARM Foundation. (2024). https://www.disarm.foundation/disinformation.

Drury, J., Reicher, S., & Stott, C. (2003). Transforming the boundaries of collective identity: From the 'local' anti-road campaign to 'global' resistance? *Social Movement Studies*, 2 (2): 191–212.

FERMI. (2024). Retrieved from Fake News Risk Mitigator: https://fighting-fake-news.eu/

FERMI. (2024). *FERMI Community Resilience and Socioeconomic Watch*. Retrieved from https://fighting-fake-news.eu/articles/fermi-community-resilience-and-socioeconomic-watch

Gelfand, M., Jackson, J., Pan, X., Nau, D., Dagher, M., & Chiu, C. (2020, April 1). Cultural and Institutional Factors Predicting the Infection Rate and Mortality Likelihood of the COVID-19 Pandemic. Available via PsyArXiv at https://psyarxiv.com/m7f8a/.

George et al. (2021).

Hameleers, M. (2023). Disinformation as a context-bound phenomenon: toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, 33(1), 1-10.

HLEG on Fake News and Online Disinformation. (2018). *A multi-dimensional approach.* European Commission.

Holvoet, M. (March 2022). International Criminal Liability for Spreading Disinformation in the Context of Mass Atrocity. *Journal of International Criminal Justice*, 223-250.

Kapiriri, L., & Ross, A. (2020). The Politics of Disease Epidemics: a Comparative Analysis of the SARS, Zika, and Ebola Outbreaks. *Global Social Welfare*, 7, 33–45.

Keijser, B., Wessels, M., & Vries, S. d. (2023). *Toekomstgerichte analyse van veiligheidsproblemen: Begrippenkader, praktijkvoorbeelden en een evaluatiekader.* Den Haag: TNO.

Keijser, Wessels, & de Vries. (2022).

Klein, A. (2012). Policing as a causal factor - A fresh view on riots and social unrest. *Safer Communities*.

Kruijver, K., Cadet, B., Finlayson, N., & Meer, S. v. (2023). *Deliverable 2.4 – Causes, contents and consequences model.* VIGILANT: VITAL INTELLIGENCE TO INVESTIGATE ILLEGAL DISINFORMATION (101073921).

Renn, O., Jovanovic, A., & Schröter, R. (2011). *Social Unrest.* Paris: OECD.

Roozenbeek, J., & Van der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of risk research*, 22(5), 570-580.

Shiffman, G. (2020, October 2). *Why Do Some Protests Turn Violent and Others Don't?* . Retrieved from https://www.lawfareblog.com/why-do-some-protests-turn-violent-and-others-dont

Stekelenburg, J. v., & Klandermans, B. (2013). The social psychology of protest. *Current Sociology*, 61(5-6), 886-905.

Sullivan, H. (2019). Sticks, stones, and broken bones: Protest violence and the state. *Journal of Conflict Resolution*, 63(3), 700-726.

Taylor, S. (2019). The Psychology of Pandemics: Preparing for the next global outbreak of infectious disease. *Newcastle-upon-Tyne: Cambridge Scholars Publishing*.

Thomas, E. (2020, October 2). *Why do protests turn violent? It's not just because people are desperate*. Retrieved from https://theconversation.com/why-do-protests-turn-violent-its-not-just-because-people-are-desperate-139968

Valentic, Z. (2024). *Deliverable 2.2 – Ethics Guidelines for PAs.* VIGILANT: VITAL INTELLIGENCE TO INVESTIGATE ILLEGAL DISINFORMATION (101073921).

10. Appendix

10.1. Effect Model Example: The Mass Gathering Model

Examples of impact types in mass gathering model (Full model shown in Figure 26)**:**

- **Economical crisis** (narratives about)**:** A narrative in the news describing a situation in which the economy of a country or region experiences a sudden downturn in its aggregate output or real gross domestic product (GDP).
- **Government budget cuts** (narratives about)**:** Messages about the narrative that government cuts spending.
- **Decline in public services** (narratives about)**:** Messages that describe backlogs, staff shortages or cuts, that could lead to health, justice and other vital systems to be less capable of functioning properly in the way they used to.
- **Political event** (narratives about)**:** An event of any kind or nature related to politics to raise intended to act as a fundraiser or political support for a political party including, but not limited to, speeches, receptions, breakfasts, luncheons, dinners, or testimonials.
- **Spread of disinformation:** Disinformation—which includes false and out-of-context information spread with the intent to deceive or mislead—is largely propagated by people looking to distort public opinion and advance particular agendas.
- **Group size:** The number of individuals within a group.
- **Increasing group sentiments:** Group sentiments refers to the moods, emotions and dispositional affects of a group of people.
- **Instigation:** Prompting or urging or encouraging someone to commit a crime or a risky activity
- **Mobilisation online:** Mass mobilization online is defined as a process that engages and motivates a wide range of partners and allies at national or local levels to raise awareness of and demand for a particular development objective through face-to-face dialogue, such as for example a mass gathering.
- **Extremist behaviour online:** Extremist behaviour is the promotion or advancement of an ideology based on violence, hatred, or intolerance, that aims to: (1) negate or destroy the fundamental rights and freedoms of others; or. (2) undermine, overturn, or replace the governments system of liberal parliamentary democracy and democratic rights.
- **Mass gathering:** A mass gathering has been defined as an occasion, either organized or spontaneous where the “number of people attending is sufficient to strain the planning and response resources of the community, city, or nation hosting the event” (WHO, 2008). These events can be planned or spontaneous, and may be as

diverse as social, religious, cultural, or sporting events or include the gathering as the result of natural disasters or conflict. Mass gatherings present their own unique challenges to public health and other risks.

- **Personal assault:** When a person directly or indirectly applies force intentionally to another person without their consent. It can also occur when a person attempts to apply such force, or threatens to do so, without the consent of the other person.

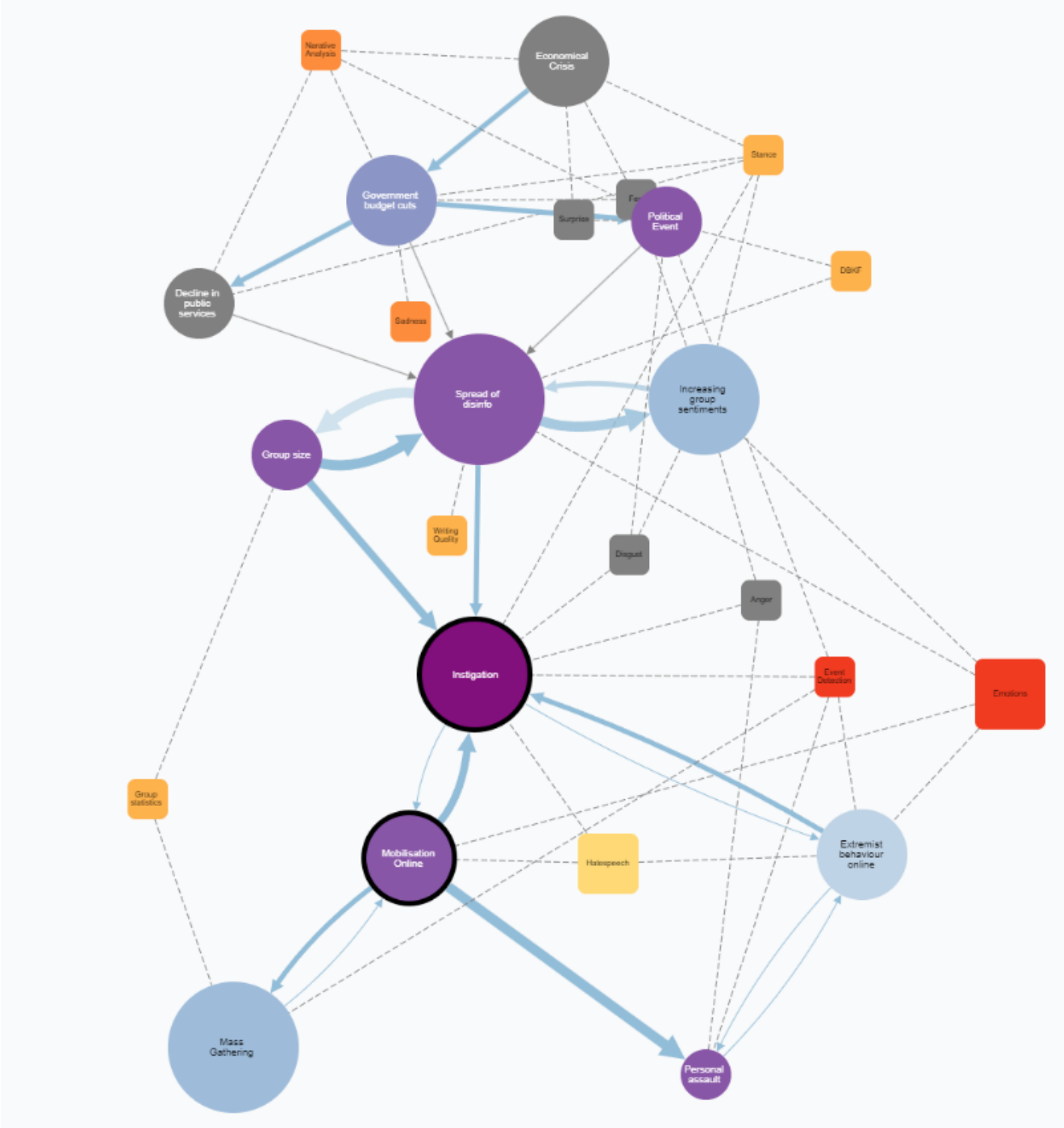


Figure 26: Visual impression of Mass gathering model with nodes (circles) and indicators (squares) connected to each other. Indicators can give indications to multiple nodes.

Escalation levels in the mass gathering model:

Level 1

- (narratives about) Economical crisis

- (narratives about) Government budget cuts
- (narratives about) Decline in public services
- (narratives about) Political event

Level 2

- Group size
- Spread of disinformation
- Increasing group sentiments

Level 3

- Instigation
- Mobilisation online

Level 4

- Extremist behaviour online
- Personal Assault
- (illegal) Mass Gathering

10.2. Effect Model Example: The Fraud Model

Another example of an effect model is the fraud model shown in Figure 27 . This model is more focussed on positive and negative feedback loops than the mass gathering model described above. The main difference between the two models is the initial approach to the design. The mass gathering model is designed from the bottom-up, starting with low impact, contextual behaviour and ramping up to more escalating and impactful behaviour. The fraud model is designed top-down, centring on the quantity of online fraud and using scientific literature to construct the key feedback loops. This partial design represents the escalation model of the fraud environment. These feedback loops

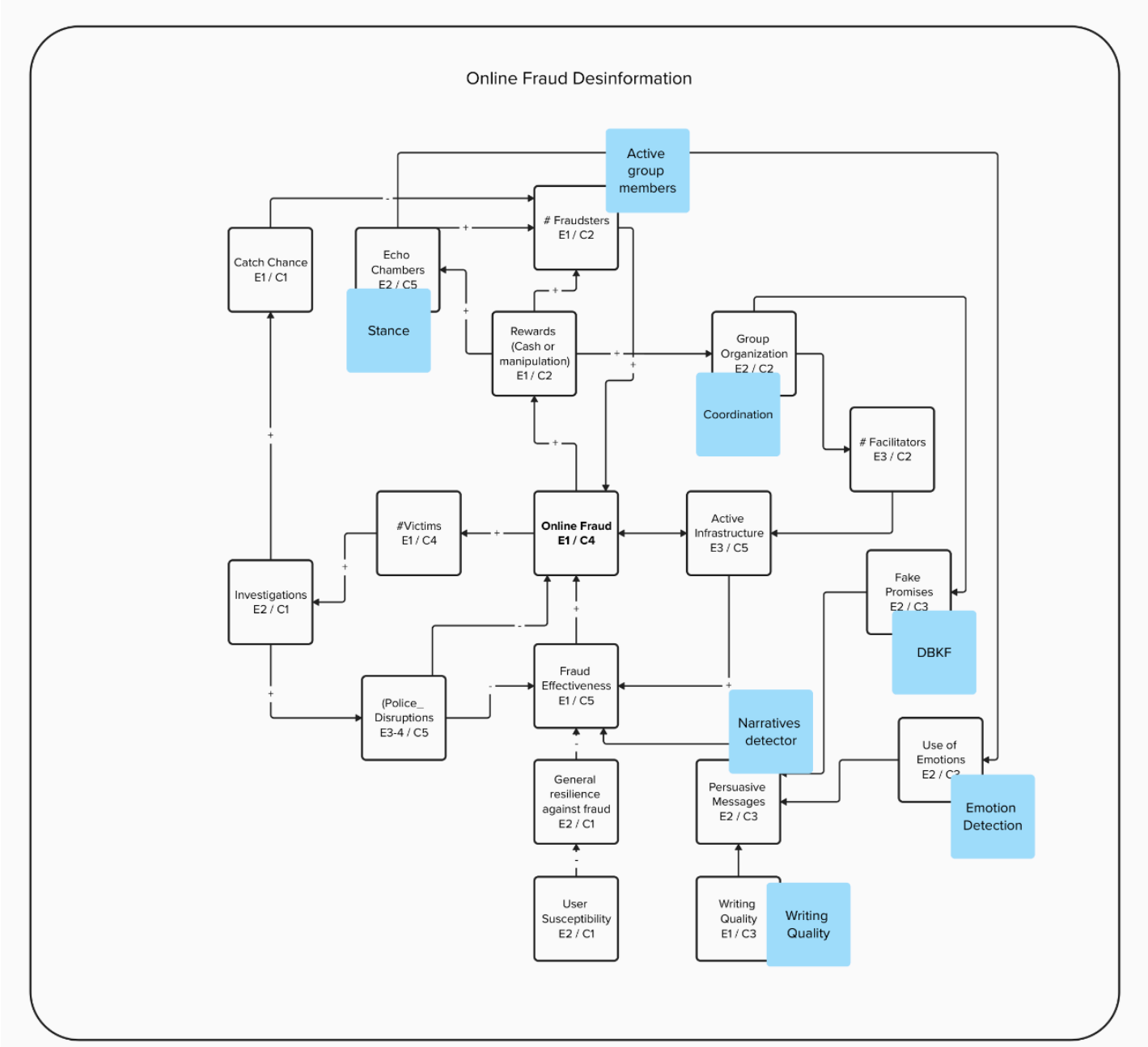


Figure 27: The fraud effect model. The connections show positive or negative correlations. In blue are the VIGILANT components that can measure the associated indicator.

are subsequently extended with disinformation indicators from the C5 model to add disinformation context and content indicators (such as writing quality).

The model consists of 18 nodes, which heavily simplifies the model while still allowing a fair amount of complexity. This fraud model is focused around one effect, with multiple model indicators, seven of which can be measured using components made in the VIGILANT program. The model is built on three main feedback loops: the number of fraudsters that flood the online environment with fake and/or misleading messages increases the presence of fraudulent messages, which increases the (likelihood of) harm done, which increases the successful examples that pull more people to become fraudsters. The measure of coordination influences the effective use and development of fraud infrastructure, which in turn increases the amount of successful fraud, gaining more resources to develop more infrastructure. The last feedback loop is the involvement of authorities and measures taken to reduce fraud. In general, the more widespread criminal fraud is in an environment, the more victims report it, which leads to a higher likelihood of police actions disrupting the fraudulent practices.

The fraud model provides a structured, qualitative framework to identify and trace the usage of disinformation within fraudulent activities.