VIGILANT

VITAL INTELLIGENCE TO INVESTIGATE ILLEGAL DISINFORMATION

Deliverable 2.3 – Ethical oversight

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: Ethical oversight
Work Package: 2 – Ethics, Disinformation and Requirements
Lead beneficiary: ALUF
Due Date: 31/10/2025
Revision Number: V1.0
Author: Elisa Orru & Zlatko Valentic
Dissemination Level: Public
Deliverable Type: Report

Overview: Deliverable D2.3 is the final ethical report of the VIGILANT project. It synthesises all ethics-related activities, assessments and mitigation measures undertaken within Task 2.3 (Ethical Oversight) from month 1 to month 36.

- **Foundation:** The analysis rests on the Ethics Framework set out in Deliverable D2.1, which defines five guiding principles—autonomy, data protection, transparency, fairness and democracy.
- **Operational guidance:** It integrates the practical guidelines for Police Authorities produced in Deliverable D2.2, turning the framework into operational rules.
- **Independent scrutiny:** All work has been continuously reviewed by the Ethics Advisory Board (EAB), whose comments and recommendations are fully reflected.

A central theme, revisited in the concluding section, is the tension between combating disinformation and safeguarding freedom of expression in police practice. The report therefore provides a comprehensive, end-to-end account of how ethical considerations have shaped the conception, design, testing and forthcoming deployment of the VIGILANT platform—while critically assessing how fundamental rights can be preserved in the fight against disinformation.

REVISION HISTORY

Version #	Implemented by	Revision Date	Description of Changes
V0.1	Zlatko Valentic	01/05/2025	Creating outline of the document
V0.2	Zlatko Valentic	07/08/2025	Completion of Part 1, 2 and 4
V0.3	Elisa Orru	13/08/2025	Completion of Part 3
V0.4	Zlatko Valentic	03/10/2025	Completion Conclusion

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

APPROVAL PROCEDURE

Version #	Deliverable Name	Reviewed by	Institution	Approval Date
V1.0	D2.3	Eva Power	TCD	23/10/2025
V1.0	D2.3	Brendan Spillane	UCD	28/10/2025

TABLE OF ACRONYMS

Acronym	Definition
AI	Artificial Intelligence
ALUF	Alberts-Ludwigs-University Freiburg
ALTAI	Assessment List for Trustworthy AI
EU	The European Union
EAB	Ethics Advisory Board
FRIA	Fundamental Rights Impact Assessment
OCR	Optical Character Recognition
PA	Police Authority
UK	United Kingdom

TABLE OF CONTENTS

1 Introduction

VIGILANT is a Europe-wide collaborative research and development project funded under HORIZON-CL3-2021-FCT-01 (Grant Agreement 101073921). It brings together research institutes, technology providers and police and security authorities from eight EU Member States in a partnership that bridges academic expertise, technical innovation and operational policing practice. The project's overarching objective is to design and deliver a modular, AI-enabled platform that supports PA's (Police Authorities) in detecting, analysing and effectively countering disinformation campaigns that threaten public security and the integrity of democratic processes. From the outset, VIGILANT has been guided by an Ethics-by-Design approach: transparency, accountability and respect for human rights are not treated as afterthoughts but are embedded in every stage of the system's design, implementation and testing.

A central challenge runs through this endeavour. Disinformation is not merely a matter of factual inaccuracy; it often overlaps, in practice, with contentious but legitimate political speech. The project must therefore reliably distinguish strategically motivated false information from genuine expressions of opinion while safeguarding fundamental rights, sustaining pluralistic debate and preserving the democratic public sphere. To meet this challenge, the ethical oversight in VIGILANT draws upon two complementary foundations.

First, the Ethics Framework developed in Deliverable 2.1 sets out five core categories—autonomy and human decision-making, data protection and privacy, transparency and interpretability, fairness, diversity and non-discrimination, and societal well-being and democratic resilience—and translates them into concrete requirements and assessment questions. Secondly, the Fundamental Rights Impact Assessment (FRIA) provides a human-rights lens, emphasising non-discrimination, the necessity and proportionality of personal-data processing, and the protection of the freedoms of expression, information, assembly and association. Taken together, these foundations ensure that the project's ethical compass is aligned both with best practice in trustworthy AI and with binding European fundamental-rights standards.

1.1 Purpose of D2.3 and its place within the overall project

Deliverable 2.3, "Ethical Oversight", marks the conclusion of Task 2.3 and constitutes the final synthesis of all ethical activities conducted from month 1 to month 36. Its purpose is threefold. It provides a transparent account of how ethical considerations have shaped the conception and development of the platform; it identifies and documents the key risks encountered, together with the measures adopted to mitigate them; and it formulates recommendations that can inform future AI-based security projects facing similar challenges.

In performing this role, D2.3 brings together the conceptual architecture established in D2.1, the operational guidance prepared for Police Authorities in D2.2, and the insights and advice of the Ethics Advisory Board (EAB). The D2.1 framework ensures that the project's ethical criteria are not merely declarative but operationalised into practical benchmarks. The D2.2 guidelines translate these benchmarks into day-to-day expectations for users who will apply the tools in real policing contexts,

thereby strengthening accountability and fostering ethical awareness beyond the development team. The EAB contributes an independent, interdisciplinary perspective that tests assumptions, questions blind spots and validates mitigation strategies. By integrating these strands into a single line of sight, D2.3 offers a coherent account of the project's ethical quality and provides a clear narrative of learning across the full lifecycle of VIGILANT.

1.1 Oversight Methodology

The methodology underpinning ethical oversight in VIGILANT rests on three interconnected pillars that together convert high-level principles into practical governance of the development process.

The first pillar is a principle-based assessment that applies, consistently and systematically, the five categories defined in D2.1 alongside the three FRIA criteria. Every component of the platform—whether a detection model, a risk-scoring module or a visualisation tool—was examined not only for AI-specific qualities such as transparency and bias, but also for its potential impact on non-discrimination, on the necessity and proportionality of personal-data use, and on freedoms of expression and association. This dual lens ensured that questions of legality and legitimacy were addressed together, and that design decisions were scrutinised for both their technical robustness and their human-rights implications.

The second pillar is a colour-coded risk matrix that communicates the outcome of ethical evaluations in an accessible, decision-oriented manner. Each tool or module is assigned a status—green, yellow, orange or red—corresponding to its ethical risk profile. Green indicates that no discernible ethical risk has been identified; yellow signals a potential risk that is manageable with appropriate safeguards; orange marks a significant risk requiring design changes and close monitoring; red denotes an unacceptable risk that demands suspension or substantial redesign. Because the matrix is applied iteratively to prototypes, updates and integrations, it functions as a continuous feedback mechanism. Development teams can see at a glance where attention is most needed, and they can track whether mitigation measures have the intended effect over time. In this way, ethical compliance becomes an integral consideration at the design stage, rather than a retrospective hurdle.

The third pillar is ongoing consultation with the Ethics Advisory Board. The Board offers external perspectives that complement the internal assessments, helps interpret ethical criteria in light of emerging project realities and flags risks that might not be captured by routine checks. Its working arrangements are being refined in consultation with the relevant partners so that its contribution can remain responsive and proportionate as the platform evolves. What matters for present purposes is that the EAB provides a structured channel for independent scrutiny and reasoned challenge, thereby reinforcing the credibility and resilience of the oversight process.

By combining these three elements—principle-based and rights-based assessment, a pragmatic and communicative risk-matrix, and independent expert oversight—the methodology ensures that ethics in VIGILANT functions not merely as a retrospective compliance exercise, but as a forward-looking steering mechanism for responsible innovation. It guides technical

development towards solutions that are both effective against disinformation and consonant with the liberal-democratic values they are intended to defend.

1.2 Structure of this Report

The document is organised into four main parts, each building on the previous section:

Part	Focus	Key content
1. Methodology	Conceptual foundation	Rationale and process by which the core ethical criteria were developed, drawing on the framework established in D2.1.
2. Tool Assessment	Practical application	Illustrative cases showing how those criteria were applied to individual VIGILANT components, using the colour-coded risk matrix.
3. Ethics Advisory Board	Governance & oversight	Composition, remit, and (evolving) working practices of the EAB, and its contribution to the ongoing ethical appraisal of the project.
4. Freedom of Expression vs. Disinformation Control	Normative reflection	A theory-informed, practice-oriented analysis of the tension between ethical guidelines and political speech in a liberal democracy, including dual-use considerations.

2 Methodology: Ethics Framework

The Ethics Framework developed in Deliverable 2.1 constitutes the normative foundation for the ethical oversight of the VIGILANT project. It is based, on the one hand, on the Ethics Guidelines for Trustworthy AI issued by the European Commission's High-Level Expert Group on Artificial Intelligence, and, on the other, on an independent and in-depth examination of the question: What, in the practical operational context of VIGILANT, should be understood as "disinformation"?

2.1 Starting Point: The EU Definition and its Limitations

As a definitional basis, the project adopted the 2018 formulation of the European Commission:

"Disinformation is false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit." (High Level Group on Fake News and Online Disinformation, 2018)

While this definition is normatively clear, its application in practice is challenging. In particular, the emphasis on ***intentional*** deception raises several problems:

- Certain forms of false or distorted information arise without direct intent to deceive, yet can have significant societal impact.
- Conversely, there is a risk that controversial or sharply worded expressions of opinion are too readily classified as disinformation, despite being protected under the fundamental right to freedom of expression.

As D2.1 stressed, any ethical assessment operates in a field of tension between the **protection of democratic discourse** and the **protection of fundamental rights**.

2.2 Differentiating Phenomena: The Scale of Information Quality

To address this tension and make it operationally manageable for VIGILANT, D2.1 developed the **Scale of Information Quality**. This classifies different phenomena according to criteria such as intention, relationship to truth, verifiability, purpose, context of use, and ethical classification. The principal forms of disinformation identified in D2.1 are:

- **Public Lie**: A deliberate, verifiable falsehood with political or societal objectives.
- **Fake News**: A mixture of true and false elements designed to manipulate perception, often emotionally charged and difficult to disprove.
- **Bullshit**: Indifference to the truth, with the primary aim of attracting attention or generating profit; often without direct political intent, yet still harmful to public discourse.
- **Propaganda & Conspiracy Theories**: Deliberate construction of alternative realities, detached from verifiable facts, with the aim of ideological conditioning and the exclusion of opposing views.

This typology enables a tailored ethical evaluation of each form, rather than applying the EU definition indiscriminately.

2.3 Political Case Studies as Benchmarks

The analysis in D2.1 was illustrated with concrete, high-profile examples that demonstrate the practical challenges of distinguishing between legitimate political communication and harmful disinformation.

One such case is the *Brexit campaign slogan "£350 million for the NHS"*, used prominently by the "Vote Leave" campaign in 2015–2016. The message, displayed alongside imagery of a bus in campaign materials, suggested that the United Kingdom was sending £350 million per week to the European Union, and that this money could instead be allocated to the National Health Service. While superficially grounded in a real financial figure, the claim was highly misleading: it did not account for the UK's rebate negotiated under EU arrangements, nor for the funds returned to the UK through EU programmes. The slogan exemplifies how strategic framing, selective omission of context, and emotionally charged national symbols can create a distorted perception without making an outright false statement. For VIGILANT, this illustrates the difficulty of classifying such material: although misleading, it operates within the grey zone between political persuasion and intentional deception, requiring careful contextual and ethical assessment.

Another example is the *US presidential campaigns of Donald Trump* in 2016 and 2020, which featured a sustained use of *public lies*, *alternative facts*, and targeted emotional appeals. In 2016, the campaign frequently amplified unverified or false claims—such as the assertion that large numbers of illegal votes were cast—while framing them in ways designed to provoke strong emotional responses, particularly fear, anger, and resentment. In 2020, similar strategies intensified, culminating in the false narrative that the presidential election had been "stolen" through widespread fraud. This narrative, widely disseminated across social media and reinforced by political allies, played a central role in mobilising supporters for the events of 6 January 2021, when the US Capitol was stormed. Here, the disinformation was not subtle: it involved verifiably false claims, sustained repetition, and an explicit political objective. Yet even in this case, classification requires more than fact-checking—it demands an understanding of the strategic intent, the emotional and political context, and the mechanisms of dissemination.

Together, these examples highlight a central finding of D2.1: the detection and classification of disinformation cannot be reduced to automated fact-verification or keyword analysis. The Brexit case shows the importance of evaluating framing, imagery, and omission of context, while the Trump campaigns demonstrate how persistent falsehoods can be weaponised through emotional and political mobilisation. Both cases confirm that effective disinformation oversight requires **political–ethical contextual analysis**, in which technical detection is combined with normative judgement to account for intention, societal impact, and the boundaries of democratic speech.

2.4 Linking to the Ethical Categories of the Framework

The **five core categories** of the VIGILANT Ethics Framework provide the structural lens through which all ethical oversight in the project is conducted:

1. **Autonomy and human agency**
2. **Privacy and data protection**
3. **Transparency and interpretability**
4. **Fairness, diversity, and non-discrimination**
5. **Societal benefit and democratic resilience**

Each category reflects a distinct but interrelated dimension of ethical risk and responsibility. Together, they define the normative standards that guide both the design and the deployment of the VIGILANT platform.

In D2.1, these categories were not treated as abstract principles alone. Instead, they were **operationalised** – translated into concrete requirements, guiding questions, and evaluative criteria – to make them usable as practical tools for assessment. This process involved identifying what each category demands in real-world terms, specifying the kinds of harms it seeks to prevent, and clarifying how compliance can be demonstrated in the context of disinformation detection.

Once operationalised, these categories were systematically mapped against the different forms of disinformation identified in the Scale of Information Quality. The aim of this mapping was to highlight how each type of information disorder engages specific ethical concerns. For example, a public lie directly undermines autonomy by distorting the factual basis for decision-making, while propaganda poses acute risks to democratic resilience by eroding trust in institutions. The result is the following overview table, which serves two purposes:

- It captures the **typical ethical impact profile** of each disinformation type across all five categories.
- It provides a **practical reference framework** for guiding both technical development and policy decisions in the VIGILANT project, ensuring that interventions address the right risks without infringing fundamental rights.

Form of Information/ Disinformation	Autonomy & human agency	Privacy & data protection	Transparency & interpretability	Fairness, diversity & non-discrimination	Societal benefit & democratic resilience
Ideal form of correct information	Supports informed decision-making by providing accurate, verifiable facts.	Respects privacy through correct and lawful use of data.	Transparent and verifiable; sources are clear and traceable.	Promotes fairness by ensuring equal access to accurate information.	Strengthens democratic processes by fostering trust and informed debate.
False News (Misinformation)	May unintentionally mislead; requires human review and timely correction to restore informed judgement.	Can inadvertently expose or misattribute personal data; prompt rectification and minimisation needed.	Error-based and typically correctable; transparency via clear sourcing and published corrections.	May reflect selection/reporting biases; mitigated through diverse sourcing and editorial checks.	Short-term confusion; long-term trust preserved if corrections and accountability occur, otherwise erosion of trust.
Public Lie	Undermines informed decision-making by distorting the factual basis.	May distort or misuse personal data to discredit groups.	Conceals intentions and obstructs traceability.	Can be targeted at specific groups, reinforcing prejudice.	Undermines trust in institutions and democratic processes.
Fake News	Influences decisions through emotionalised and partly false depictions; hinders critical assessment.	Often linked to personalised data for targeted impact.	Mix of truth and falsehood makes verification difficult.	Often contains discriminatory undertones or stereotypes.	Reinforces polarisation; hinders fact-based debate.
Bullshit	Fosters disinterest in truth; weakens rational judgement.	Uses mass-harvested data for click generation, often without personal focus.	Lacks sources or justification; truthfulness irrelevant.	Can unintentionally reinforce prejudice.	Promotes cynicism towards public discourse.
Propaganda & Conspiracy Theories	Replaces individual judgement with ideological narratives.	Exploits personal data for manipulation/surveillance.	Employs unverifiable narratives; resists critique.	Systematically fosters enemy images; excludes groups from discourse.	Aims to erode democratic values and institutions.

2.5 Ethical Depth and the Dual Mandate of VIGILANT

Integrating this typology into the Ethics Framework makes clear that VIGILANT operates under a *dual mandate*:

1. To *detect and reduce harmful disinformation* that undermines democratic processes and erodes social cohesion.
2. To *safeguard freedom of expression* and avoid misclassifications that could suppress legitimate political speech.

Balancing these two objectives is a constant ethical challenge. The risk of overreach is particularly acute in borderline cases – for example, pointed political criticism, satirical exaggeration, or controversial yet fact-based argumentation. In such cases, the difference between harmful manipulation and protected speech is not a matter of mere technical analysis; it requires nuanced human judgement informed by legal standards, political context, and ethical reasoning.

The methodology developed in D2.1 reflects this understanding. It evaluates disinformation not solely through technical detection mechanisms, but also within its *socio-political context*, recognising that the same content may be harmful in one setting but legitimate in another. This approach is therefore an indispensable element of ethical oversight in D2.3.

To ensure that such sensitivity is maintained in practice, *guiding questions* have been formulated for each of the five core categories of the Ethics Framework. These are not numerical thresholds – indeed, the ethical risks involved can rarely be reduced to quantifiable measures. Rather, they are *qualitative prompts* designed to structure human deliberation, which may, where appropriate, be translated into measurable indicators for monitoring purposes. Even then, it must be recognised that the true danger posed by disinformation can never be fully "measured" in any definitive sense: *it will always ultimately depend on the exercise of human judgement*.

2.5.1 Guiding Questions per Category

1. Autonomy and human agency

- Does the tool support users' decision-making, or does it risk uncritical acceptance of its outputs?
- Is it made explicit that final responsibility for assessment always rests with a human operator?
- Are mechanisms in place to challenge or override system decisions?
- Are outputs designed to foster critical thinking rather than replace it?

2. Privacy and data protection

- Is the data that is processed strictly necessary for the specified purpose?
- Is all data processing compliant with applicable data protection laws (e.g. GDPR)?
- Is all the personal data safeguarded against misuse, re-identification, or repurposing?
- Are there clear processes for deleting or anonymising sensitive data?

3. Transparency and interpretability

- Can users and oversight bodies understand how the tool arrived at a given classification?
- Are data sources, methodologies, and model limitations documented and accessible?
- Are examples or explanations provided to clarify the tool's assessments?
- Are warnings issued in cases of uncertainty or potential misclassification?

4. Fairness, diversity, and non-discrimination

- Is it assessed whether the tool disproportionately flags certain groups as sources of disinformation?
- Are training data and evaluation standards selected to minimise discriminatory bias?
- Are different perspectives and contexts taken into account to avoid misclassification?
- Are there processes in place to correct identified biases?

5. Societal benefit and democratic resilience

- Does the tool contribute to protecting democratic processes and strengthening both public and institutional trust in information systems?
- Does it ensure that legitimate expressions of opinion are not wrongly labelled as disinformation, including in cases assessed or monitored by Police Authorities?
- Is information communicated in ways that build trust — not only among citizens, but also between Police Authorities and the public they serve?
- Are safeguards in place to prevent misuse for politically motivated purposes and to ensure proportionality and accountability in law enforcement contexts?

Through this qualitative operationalisation, the Ethics Framework is not merely a set of normative principles but also a *practical instrument* for ongoing ethical oversight in the VIGILANT project. The guiding questions serve as a foundation for structured reviews, as a common reference point for dialogue between technical and ethical partners, and as a basis for deriving quantitative indicators in deployment contexts – while always recognising that ethical assessment remains, at its core, a matter of human discernment.

3 Application Of The Framework To The Vigilant Tools

3.1 Ethical Guidelines and Tool Evaluation

The ethical assessment of the VIGILANT platform was conducted using a combined evaluation framework, bringing together two established approaches: the *Assessment List for Trustworthy Artificial Intelligence (ALTAI)*, with its seven requirements — human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well-being, and accountability — and the three criteria of the *Fundamental Rights Impact Assessment (FRIA)* — non-discrimination, protection of personal data (necessity and proportionality), and freedom of expression, information, assembly and association.

By applying both frameworks in parallel, each tool was assessed not only against general ethical standards for AI but also for its specific implications for fundamental rights, which are particularly relevant in a law enforcement context. The assessment used a four-tier colour-coded risk matrix (Green = no risk, Yellow = manageable with safeguards, Orange = significant risk requiring mitigation, Red = unacceptable).

Colour	Meaning
Green	No discernible ethical risk
Yellow	Potential risk, manageable with suitable measures
Orange	Clear risk, requiring heightened attention and corrective action
Red	Unacceptable risk, necessitating project suspension or major redesign

3.2 Cross-cutting Risk Areas

A number of risk themes recurred across multiple modules:

- **Transparency and interpretability**: Complex models, especially in network analysis, often produce results that are difficult for non-technical users to understand without additional context. ***Mitigation***: Explainability features ("Why was this flagged?"), clear documentation, mandatory operator training.
- **Fairness and non-discrimination**: Training data may contain unintended biases that disproportionately flag certain groups or linguistic patterns. ***Mitigation***: Regular bias audits, bias detection tools, manual review of sensitive classifications.
- **Privacy and proportionality**: Higher risk where modules process identifiable content or metadata. ***Mitigation***: Default anonymisation, strict access controls, short retention periods.
- **Freedom of expression and democratic resilience**: Risk of misclassifying legitimate political speech, satire or fact-based controversial statements as disinformation. ***Mitigation***: Human review in high-risk cases, context-sensitive analysis, transparent appeal mechanisms.

- **Human oversight and accountability**: Lack of clearly defined points for mandatory human intervention. ***Mitigation***: Defined "human-in-the-loop" checkpoints, full logging of decisions.

3.3 Tool-specific Risk Profiles

Writing Quality

This module presents a higher risk of indirect discrimination, particularly against non-standard dialects or spelling errors, which could be misinterpreted as indicators of low credibility.

- *Main concern*: Misclassification based on linguistic variation.
- *Recommendation*: Clear evaluation criteria, bias testing, human review before action.

EDNA Event Detection

Automatic event detection can produce false alerts where data is incomplete or inaccurate, potentially triggering unnecessary interventions.

- *Main concern*: False positives in event identification.
- *Recommendation*: Cross-check with multiple data sources, disclose methodological limitations.

Multilingual OCR

Challenges arise with languages and scripts for which training data is limited, leading to recognition errors and biased outputs.

- *Main concern*: Language and script bias.
- *Recommendation*: Language coverage tests, manual verification, privacy-by-design.

Coordination Detection

This module carries particularly high risks for freedom of communication and privacy, as it may involve the analysis of large volumes of communications data.

- *Main concern*: Overreach into the activities of uninvolved parties.
- *Recommendation*: Strict usage boundaries, authorisation processes, clear data deletion protocols.

Stance / Hate Speech / Emotion Analysis

These modules are prone to misclassification, especially where cultural differences in expression are not adequately considered.

- *Main concern*: Loss of context leading to premature or inaccurate labelling.
- *Recommendation*: Lower automatic thresholds, rights of appeal, stakeholder input.

Impact Analysis

While useful as an early warning system, there is a risk of overinterpreting quantitative measures such as the "Narrative Matching Score".

- *Main concern*: Overestimation of accuracy.
- *Recommendation*: Keep outputs advisory rather than determinative, provide clear explanations of results.

3.4 Guidelines for Police Authorities

The findings from the combined ALTAI–FRIA evaluation can be translated into a clear operational framework for Police Authorities. Above all, in high-risk contexts, decisions based on outputs from the VIGILANT tools should never be taken in a fully automated manner. Human judgement must remain the decisive element, ensuring that the outputs of automated systems are balanced against contextual understanding, operational priorities, and the protection of fundamental rights.

Strict adherence to data protection principles is equally non-negotiable. Data minimisation, anonymisation or pseudonymisation, and the conduct of Data Protection Impact Assessments (DPIAs) should be standard practice, particularly when handling sensitive or personally identifiable data. The functioning of each tool — including its limitations and the criteria used for flagging content — must be made transparent both internally and to relevant external stakeholders. Biases in datasets and algorithms must be actively identified and mitigated, while freedom of expression should be safeguarded through precise thresholds, context-sensitive human review, and accessible redress mechanisms. Finally, accountability structures must be explicit: responsibilities for model maintenance, operational decisions, error correction, and public complaints must be clearly defined, supported by continuous monitoring and regular reassessment of risk levels.

Within this overarching framework, eight core guidelines provide a structured approach for Police Authorities seeking to deploy VIGILANT responsibly:

1. Human oversight should always be prioritised. Human decision-making remains indispensable for counterbalancing the limitations of AI systems, particularly in complex or high-stakes situations. Scenarios in which human review is mandatory — such as decisions affecting personal liberty or carrying legal consequences (e.g., arrests, searches) — should be defined in advance. Personnel must be trained to interpret AI outputs critically and to override them where appropriate.

2. Continuous monitoring and evaluation are essential to safeguard both operational effectiveness and ethical integrity. Real-time monitoring of system performance should be complemented by regular reviews conducted by both internal and external experts. Automated anomaly detection should be in place to identify and flag unusual system behaviour.
3. Transparency and accountability must be ensured. Public trust depends on open communication about how AI is used, what data it processes, and the rationale behind its outputs. This information should be made publicly available in clear and accessible language and kept up to date.
4. Privacy and responsible data governance are fundamental. Police Authorities should apply strict data protection protocols — including encryption, access controls, and anonymisation where appropriate — and conduct regular reviews to ensure compliance with relevant legislation.
5. Ethical awareness and capacity-building should be embedded in organisational practice. Ongoing training programmes should address AI ethics, bias mitigation, and the relevant legal implications, making use of practical case studies and realistic simulations to prepare personnel for real-world challenges.
6. Stakeholder dialogue must be actively facilitated. Engaging with citizens, civil society organisations, and academic experts in regular consultative forums fosters democratic legitimacy and ensures context-sensitive governance.
7. Accessible review mechanisms should be implemented. Clear, transparent processes for lodging complaints or appeals against AI-assisted decisions must be publicly communicated and operate independently from the decision-makers.
8. Collaboration and knowledge exchange should be promoted. Building networks with national and international partners enables Police Authorities to share best practices, harmonise ethical standards, and jointly develop innovative approaches to responsible AI deployment.

3.5 Concluding Analysis

These guidelines underline that the ethical deployment of AI cannot be reduced to mere technical compliance. It requires the sustained integration of human oversight, transparency, and accountability into everyday operational practice. They address both systemic requirements — such as governance, training, and stakeholder engagement — and safeguards specific to AI-assisted disinformation detection, including bias prevention, proportionate data processing, and the protection of freedom of expression. Implemented consistently, they can ensure that the VIGILANT platform serves not only as an effective operational tool, but also as a mechanism for strengthening democratic resilience and public trust in law enforcement.

4 Ethics Advisory Board (EAB)

4.1 Structure and Composition

The Ethics Advisory Board (EAB) was established in the third month of the project and remained active until its conclusion. It brought together experts from a range of fields, including ethics, data protection law, policing, civil society, and technology ethics. The EAB was constituted by two members recruited among experts external to the project and one internal member (Elisa Orrù, Freiburg University) with the role of supporting the project coordination in involving the external members into the project activities.

The two external experts nominated at the beginning of the project are Maria Grazia Porcedda, Assistant Professor of IT Law at Trinity College Dublin with a research focus on cybersecurity, privacy and data protection in EU law, and Marina Markellou, Assistant Professor of law at the University of Groningen and Intellectual Property Law Attorney. In addition to the feedback provided by the EAB members, VIGILANT also consulted other ethics external experts on specific topics, including Michael Kühler, Professor of Applied Ethics in Social Responsibility at the University of Applied Sciences in Dortmund and Joanna Rozynska, Assistant Professor of Ethics at the University of Warsaw.

4.2 Mandate and Responsibilities

The Ethics Advisory Board served as an independent body providing continuous ethical guidance throughout the development and implementation of the VIGILANT platform. Its core responsibilities included:

- **Ethical consultation and review** on system design choices, especially in relation to human rights and democratic values.
- **Risk assessment** of technical components, with particular attention to dual-use potential, data sensitivity, and speech-related implications.

The Board acted as a critical sounding board for unresolved questions, helped interpret ethical criteria in light of emerging project realities, and ensured that fundamental rights were given sustained consideration across all phases of development.

4.3 Activities and feedback

The core members of the EAB were invited and attended either in presence or remotely the Project's plenary meetings, including the VIGILANT Kick-Off Meeting in Dublin in November 2022 and the Bratislava General Assembly in September 2023. Additionally, ad-hoc meetings and consultations were organised either with the core members or with the additional members to receive feedback on specific topics or on the VIGILANT demonstrators.

Topics on which advice from the EAB was sought during the early phase of the project included:

- the use of data during the project, especially for training purposes
- the definition of disinformation and the tensions with freedom of expression
- the targeted vs. mass-search functionalities the VIGILANT platform, especially regarding the data sourcing from social media platforms
- The identification of adequate standards for cybersecurity, considering the lack of all-encompassing and consistent standards.

As the project evolved, we were able to ask the external EAB members to provide feedback on the VIGILANT demonstrations. Feedback received regarding the development of VIGILANT included the following issues:

- Technical challenges for automatic recognition of disinformation, especially when innocuous symbols are used for hatful purposes
- The risks of misuse of VIGILANT for mass surveillance
- The possible resistance from the side of LEAs to take into account ethics aspects, as these can be seen as restricting security
- The risk of misuse of VIGILANT by malign actors

The discussions with the EAB members and their feedback was continuously integrated into the VIGILANT ethics work and reflected into the Ethics deliverable. Feedback received during the first phase of the project significantly contributed to the Ethics-by-design approach of VIGILANT (see section 2.3 of this deliverable), while the input received during later stages of the project also informed the guidelines for police authorities (see section 2.4 of this deliverable) and resonates with the concluding considerations on the tension between counteracting disinformation and protecting free speech that are at the centre of the next section.

5 Tension: Ethical Guidelines And The Democratic Public Sphere

5.1 The Fundamental Tension: Freedom vs. Control

The operational deployment of VIGILANT within Police environments sits at the heart of a democratic tension: on the one hand, protecting institutions, elections, public safety and opinion-formation from targeted manipulation; on the other, safeguarding freedom of expression as a cornerstone of democratic life. This tension is not incidental but inherent to any system that assesses the truthfulness, provenance or societal impact of communicative content.

In VIGILANT's case, algorithmic classifications and automated responses (such as de-prioritisation, labelling or removal suggestions) must be configured so as not to narrow, even inadvertently, the legitimate space for democratic communication.

5.2 Matrix of Tensions

As part of Task 2.3, the following matrix was developed and applied to VIGILANT's operational design. It identifies dimensions where democratic values and technical control mechanisms intersect in ways that require careful balancing:

Criterion	Democratic Public Sphere	Controlling System (Techno-Ethical Regulation)
Diversity of Discourse	Pluralism welcomed	Heterogeneity complicates automatic detection; parameter tuning to avoid overfitting to dominant speech patterns
Ambiguity & Satire	Part of expressive culture	Risk of false positives; classifier thresholds adapted and subject to human review
Critique & Opposition	Democratically legitimate	Safeguards in place to prevent misclassification as 'extreme' or 'suspicious'
Spontaneous Online Dynamics	Element of a vibrant public sphere	Continuous model monitoring to limit overreach into unpredictable debate patterns

5.3 Dual-Use Risks and Political Instrumentalisation

A specific risk arises from the potential for dual-use: tools initially developed to counter manipulative campaigns can – in the absence of sufficient ethical safeguards – be repurposed for political influence. This risk is particularly pronounced in authoritarian contexts or within platforms used in systematically partisan ways. VIGILANT has therefore deliberately implemented a governance structure that includes regular audits, transparency measures and external oversight.

5.4 Supplementary Ethical Evaluation – Freedom of Expression as a Benchmark

Despite all efforts to align the VIGILANT platform with ethical principles, a core tension remains between the application of ethical guidelines (e.g. in the fight against disinformation) and the fundamental right to freedom of expression. This tension is

particularly salient in the context of algorithmic detection, classification, and potential intervention in relation to content labelled as "disinformation".

The ethical advisory work throughout the project has highlighted several key insights:

- Ethical guidelines are essential to tether technical systems to democratic values – but they must not become a normative overreach into the realm of legitimate political speech.
- There is a significant risk that technical and normative criteria – particularly those aimed at truthfulness or social harmony – may inadvertently capture dissenting, critical or marginalised voices, thereby narrowing the space for public debate.
- Freedom of expression, as enshrined in Article 11 of the EU Charter of Fundamental Rights and Article 10 of the European Convention on Human Rights, must therefore serve as a robust benchmark – even when ethical interventions are otherwise well-founded.

5.5 Recommendations: Principle-Led Balancing Procedures

To ensure that VIGILANT's operational roll-out strengthens democratic resilience while avoiding the suppression of legitimate speech, the following three interlinked recommendations are embedded in deployment protocols:

#	Recommendation
1. Build Theoretical Awareness	Introduce a structured programme of conceptual training for all stakeholders—engineers, data scientists, police officers and policy-makers alike. The curriculum should cover: - the democratic role of free expression and open debate; - the epistemic and societal harms posed by organised disinformation; - the ethical principles (autonomy, privacy, fairness, transparency, democracy) that underpin VIGILANT's design. This theoretical grounding equips practitioners to understand why the fight against disinformation matters—beyond operational imperatives—and helps them internalise the normative limits within which technological counter-measures must remain.
2. Institutionalise a Pluralistic Ethics Board	Maintain (and, where possible, expand) a pluralistically composed ethics body. "Pluralistic" here means drawing expertise from: - technical disciplines (AI, cybersecurity), - social sciences (psychology, sociology, media studies), - humanities (history, philosophy, ethics), - frontline practice (law-enforcement, journalism, civil society). Such breadth ensures that recommendations reflect diverse methodological outlooks and lived experiences, producing balanced and context-sensitive guidance rather than silo-specific prescriptions.

#	Recommendation
3. Mandatory Proportionality Review	Embed a formal proportionality test into every planned intervention—mirroring the legal four-step logic (legitimate aim, suitability, necessity, balance of interests). The review should be documented, repeatable, and auditable, providing a transparent trail of how speech-related risks were weighed against security objectives.

Lessons from VIGILANT: Project experience shows that embedding these safeguards in the operational framework is just as critical as the technical accuracy of the tool. This insight closes the ethical reflection of Chapter IV: effective counter-disinformation work depends not only on algorithmic precision but on sustained awareness of democratic implications. These lessons prepare the ground for the concluding reflection on human judgment as the ethical core of VIGILANT.

6 Conclusion: Human Judgment as the Ethical Core of VIGILANT

The ethical oversight undertaken within Task 2.3 leads to a central insight that reaches beyond questions of technical governance: it is not only important to affirm *that* human beings must retain the final decision, but to understand *why* this is the case. The VIGILANT project has demonstrated that human judgement is not a supplementary safeguard, but a legal and ethical necessity—the condition under which AI-assisted policing can remain compliant with fundamental rights and democratically legitimate.

Artificial intelligence, by its very structure, operates on statistical regularities drawn from the past. Its analyses can reproduce existing biases and systemic asymmetries, and it lacks the contextual awareness required to weigh proportionality, intention and consequence. It can compute correlations, but it cannot perceive meaning. Automated decisions—particularly in domains that affect rights and freedoms—therefore cannot satisfy the requirements of accountability, legality and proportionality established in European fundamental-rights law. These obligations presuppose an agent capable of moral and legal reasoning: a human being.

From both ethical and operational perspectives, this insight became tangible throughout the project's engagement with Police Authorities. Through workshops, presentations and field discussions—most notably during the training session in Münster—officers were not merely reminded that human oversight is mandatory, but were encouraged to reflect on *why* this responsibility cannot be delegated to algorithms. This process of awareness-building proved to be one of the most significant outcomes of Task 2.3. It translated abstract ethical principles into a concrete understanding of professional responsibility and legal accountability within AI-supported environments.

This insight directly continues the reflection of Chapter IV. Every system designed to counter disinformation inevitably operates within a field of tension between freedom and control. Technical interventions intended to protect democracy may, if applied without discretion, restrict the very communicative freedoms they are meant to defend. The ethical framework developed within VIGILANT—combining principle-based assessment, iterative risk matrices and independent oversight—ensures that this

tension remains visible and manageable. It institutionalises proportionality, pluralistic review and theoretical awareness as structural elements of responsible innovation.

Seen in this light, the project experience confirms three overarching lessons:

1. Human discernment is indispensable for maintaining the balance between operational effectiveness and fundamental rights.
2. Ethical governance must be embedded in organisational practice, not added post-factum as a matter of compliance.
3. Democratic awareness is a condition of professional competence: only when operators understand the societal meaning of their actions can AI tools be used legitimately.

Accordingly, VIGILANT's contribution is twofold. On the operational level, it delivers a modular, ethically validated platform that enhances the analytical capacity of Police Authorities. On the normative level, it strengthens the democratic and legal foundations of such work by clarifying *why* human decision-making is indispensable: because only human agents can interpret context, balance rights, and assume responsibility for outcomes.

In conclusion, VIGILANT not only improves the technical detection of disinformation but reinforces the ethical and legal awareness that must accompany its use. By embedding this awareness among Police Authorities, the project ensures that AI serves democracy not as a substitute for judgement, but as a tool guided by it. The human being remains the interpretive horizon of every decision—the point at which information becomes understanding, and control becomes responsibility.