D2.3 – Hatedemics methodology v.2

HATEDEMICS is funded by the European Union
(CERV-2023-CHAR-LITI-101143249)

Project Information

Grant Agreement	101143249
Name	Hampering hate speech and disinformation through AI-based technologies to prevent and combat polarisation and the spread of racist, xenophobic and intolerant speech and conspiracy theories
Acronym	HATEDEMICS
Call	CERV-2023-CHAR-LITI-SPEECH
Topic	Topic 4 - Protecting EU values and rights by combating hate speech and hate crime
Consortium Coordinator	Fondazione Bruno Kessler (FBK)
Project duration	24 months
Project start date	1 April 2024
Project closing date	31 March 2026
Coordinator	Marco Guerini

Document Information

Workpackage	WP2
Deliverable	D2.3 Hatedemics methodology – v2
Deliverable Lead	FUNDEA
Dissemination Level	PUBLIC
Type	REPORT
Date due for submission	31/01/2026
Date submitted	11/02/2026
Authors	F. Javier Montilla-Aguilera (FUNDEA) José Luis Salido-Medina (FUNDEA) Diego de Haro (CENTRA) Helena Bonaldi (FBK)
Editor(s)	F. Javier Montilla-Aguilera (FUNDEA) José Luis Salido-Medina (FUNDEA)
Reviewers	Rubén Martín (CENTRA) Eva Cataño (CENTRA) Ana Velasco (MALDITA)

Revision History

Version	**Date**	**Author**	**Comments**
v0.1	02/02/2026	F. Javier Montilla-Aguilera (FUNDEA) & José Luis Salido-Medina (FUNDEA)	First draft of the deliverable

v0.2	09/02/2026	Ana Velasco (MALDITA) & Rubén Martín (CENTRA) & Eva Cataño (CENTRA)	Reviewers'evaluation
V0.3	10/02/2026	F. Javier Montilla-Aguilera (FUNDEA) & José Luis Salido-Medina (FUNDEA)	Revised version after addressing the comments of the reviewers, ready for final check
V1.0	11/02/2026	Pamela Forner	Quality check and Final version of the deliverable

EXECUTIVE SUMMARY

Deliverable 2.3 (D2.3) concludes Work Package 2 of the HATEDEMICS project by consolidating the methodological learnings throughout the entire project. Its main objective is to translate research evidence and user feedback into operational guidance for the design, piloting, educational use and long-term sustainability of the HATEDEMICS platform.

Building on the desk research conducted in D2.1 and the qualitative empirical work of D2.2, this deliverable represents the transition from understanding the problem space of hate speech, disinformation and conspiracy narratives to defining how an AI-based system should responsibly operate within that space. It integrates insights from focus groups, interviews, co-creation workshops, pilot activities and UX testing carried out across different national contexts, ensuring that requirements are grounded in real-world practices rather than abstract assumptions.

The platform is conceived not as a neutral technological artefact, but as a system embedded in social practices, ethical norms and educational settings. This perspective is combined with a strong ethical AI framework aligned with EU principles on transparency, explainability, human oversight and fundamental rights, as well as with a critical pedagogical orientation that prioritises reflection, media literacy and civic engagement over automation or sanctioning.

Empirical evidence consistently shows that users value AI tools that support interpretation and learning rather than deliver definitive judgments. Across countries and professional profiles, stakeholders emphasised the need for contextualisation, explainable outputs and clear communication of uncertainty. Concerns about bias, misuse, over-automation and fear amplification emerged as recurring themes, highlighting structural tensions such as automation versus human accountability, analytical depth versus cognitive load, and standardisation versus contextual sensitivity.

In response, D2.3 defines socio-technical requirements across the core components of the HATEDEMICS platform, including detection, risk estimation, AI-assisted counter-narratives and educational integration. Detection and analytical outputs are framed as interpretative aids rather than authoritative classifications. Risk indicators are designed as contextual and educational signals, with explicit uncertainty and non-alarmist visualisation. Counter-narratives are conceived as human-in-the-loop processes, where AI provides support while preserving human agency, editorial control and ethical responsibility. Educational use is fully aligned with training toolkits, guided workflows and age-appropriate safeguards.

By consolidating these requirements, D2.3 provides a shared reference framework for subsequent technical development, piloting, impact assessment and sustainability planning. It ensures that innovation within the HATEDEMICS project remains firmly anchored in social realities, ethical principles and educational objectives, supporting the broader goal of fostering critical, inclusive and resilient digital citizenship in line with European values.

Table of Contents

List of figures

List of tables

Acronyms

ABBREVIATED	EXTENDED
WP	Work Package
AI	Artificial Intelligence
NGO	Non-Governmental Organization
CSO	Civil Society Organizations
LLM	Large Language Models
STS	Science Technology and Society
SNA	Social Network Analysis
EU	European Union
HITL	Human-in-the-Loop
IFCN	Fact Checking Network
DSA	Digital Service Act
GDPR	General Data Protection Regulation
LEA	Local Educational Agency

1 INTRODUCTION

Deliverable 2.3 constitutes the consolidation and closure of Work Package 2 (WP2) of the HATEDEMICS project. Its main purpose is to translate the conceptual, empirical and participatory insights generated during the first stages of the project into a coherent and operational set of methodological steps that guide the development, piloting, educational deployment and long-term sustainability of the HATEDEMICS platform.

Within the internal logic of WP2, D2.3 builds directly upon the results of D2.1 and D2.2, functioning as the bridge between analytical research and applied system design. While D2.1 provided an extensive conceptual and contextual mapping of hate speech, disinformation and conspiracy theories, together with an overview of existing technological, legal and policy approaches, D2.2 complemented this framework with empirical qualitative evidence gathered through focus groups and in-depth interviews with key stakeholders across different national contexts. These earlier deliverables identified not only shared understandings and best practices, but also persistent ambiguities, ethical concerns and practical tensions related to the use of AI-based tools in sensitive social and educational domains. D2.3 articulates how abstract concepts, stakeholder expectations and observed user behaviours can be translated into concrete requirements shaping the functionalities, governance mechanisms and pedagogical orientation of the HATEDEMICS platform. In this sense, D2.3 represents the transition from understanding the problem space to defining how the system should behave within that space.

In addition, this deliverable integrates key insights emerging from WP4 co-creation activities and Phase 1 pilot implementations, as well as from the development of the HATEDEMICS educational toolkit. Feedback collected during workshops and national pilots provided valuable evidence on how different user groups—such as NGO practitioners, journalists, educators and young people—interact with the platform in practice, how they interpret analytical outputs, and which design features support or hinder meaningful engagement. These insights are essential to ensure that socio-technical requirements are grounded not only in theoretical assumptions, but also in real-world usage scenarios.
A core objective of D2.3 is to demonstrate how the knowledge generated within WP2 translates into actionable socio-technical requirements across four key dimensions of the HATEDEMICS system: detection, risk estimation, human-in-the-loop counter-narratives, and educational components.

First, insights from D2.1 and D2.2 highlighted the conceptual and normative complexity of defining hate speech, disinformation and related phenomena. Stakeholders consistently emphasised that rigid or purely technical definitions risk oversimplifying social dynamics and undermining trust in AI-based systems. As a result, the socio-technical requirements articulated in this deliverable prioritise contextualisation, transparency and explainability in detection mechanisms, avoiding binary classifications and emphasising interpretative support for users.
Second, with regard to risk estimation, WP2 findings underscored the need to distinguish between analytical indicators and normative judgements. Stakeholders expressed concerns about the potential misuse of risk scores if they are perceived as definitive or punitive. Consequently, D2.3 defines requirements for the HATEDEMICS Risk component as a

contextual and educational tool, designed to support awareness and strategic reflection rather than automated decision-making. Requirements therefore address issues such as confidence levels, visualisation clarity, and adaptation of outputs to different user profiles.
Third, the qualitative research and pilot results strongly informed the design of human-in-the-loop counter-narrative functionalities. Both professionals and educators stressed that counter-speech cannot be fully automated without losing credibility, ethical sensitivity and contextual relevance. The socio-technical requirements developed in this deliverable therefore frame AI-supported counter-narratives as assistive mechanisms, embedded within workflows that preserve human agency, editorial control and reflexivity. This approach directly aligns with the values and safeguards identified during WP2 research.

Finally, WP2 and WP4 jointly highlighted the importance of educational integration as a defining feature of the HATEDEMICS project. Beyond professional use, the platform is intended to function as a learning environment supporting media literacy, critical thinking and civic engagement. D2.3 therefore articulates requirements that ensure coherence between platform functionalities and the educational toolkit, including accessibility, pedagogical framing, age-appropriate safeguards and alignment with non-formal and formal learning contexts.
By consolidating socio-technical requirements across these dimensions, D2.3 plays a strategic role in ensuring alignment between research, technology and education within the HATEDEMICS project. It provides a common reference point for WP3 (technical development), WP4 (piloting and training), and WP6 (impact assessment and sustainability), supporting an iterative and reflexive development process.
In this way, D2.3 establishes a robust foundation for the second phase of the project, ensuring that technological innovation remains firmly anchored in social realities, ethical principles and educational objectives, in line with the broader goals of the Citizens, Equality, Rights and Values (CERV) program.

2 RATIONALE AND THEORETICAL FRAMEWORK

The development of artificial intelligence–based tools for the educational field poses a range of challenges that go beyond purely technical considerations. In the case of a project such as HATEDEMICS, which involves the use of a web application that in turn employs a large language model (LLM) to teach young people how to identify and counter hate speech, it is essential to adopt a sociotechnical approach that simultaneously integrates the technological, pedagogical, and ethical aspects of design. This approach makes it possible to understand the application not merely as a piece of software, but as an artefact that mediates social practices, values, and forms of learning.

2.1 Sociotechnical Perspective: Technology as Social Practice

Science, Technology, and Society (STS) studies provide a particularly suitable framework for analysing the interactions between the social and technical components of innovation. From this perspective, technology is not a neutral element, but rather the outcome of a web of decisions, negotiations, and values (Latour, 1987; Callon, 1986). As Winner (1980) argues, artefacts have politics: their design and modes of use embody worldviews, distribute power, and shape behaviour.

In this sense, all the components of the HATEDEMICS project (the chatbot, the hate and disinformation diffusion network, and the educational activities) constitute actors within a sociotechnical network that includes young people, teachers, experts, designers, and public policymakers. Each interaction with the system contributes to the production of specific meanings and learning dynamics, which makes it necessary to understand its operation in terms of situated action (Suchman, 2007): the actual uses of technology depend on the social, cultural, and emotional contexts in which it is embedded.

Adopting a sociotechnical approach therefore entails defining the application's requirements on the basis of a contextualised understanding of users' practices and expectations, as well as of the values that should guide its design: transparency, inclusion, safety, and trust.

2.2 Ethics and Governance of Artificial Intelligence

The growing adoption of AI systems in education requires a robust ethical framework. Floridi et al. (2018; Floridi & Cowls, 2019) propose the "AI4People Framework," which establishes five fundamental principles: beneficence, non-maleficence, autonomy, justice, and explicability. These principles can be translated into operational system requirements, such as ensuring that the LLM provides pedagogically constructive responses (beneficence), avoids the reproduction of biases or stereotypes (non-maleficence), respects the agency of young users (autonomy), promotes equity and diversity (justice), and, finally—and probably the most important addition to the classic bioethical principles in the context of projects involving large language models for socio-educational purposes—offers intelligible and understandable functioning (explicability).

Moreover, from the perspective of value-sensitive design (Brey, 2012), technological development should integrate human values from its earliest stages, involving future users and stakeholders in the design process. This principle aligns with the recommendations of the European Union's AI Act (2024) and its Digital Education Action Plan (2020), which emphasise the importance of social participation, accountability, and human oversight in systems with

high educational or social impact. Within this framework, the ethical dimension is not a subsequent add-on to the HATEDEMICS project, but rather a constitutive element of the system's foundational development.

2.3 Critical Education and Digital Literacy

From the perspective of critical pedagogy, education is not understood as the transmission of knowledge, but as a dialogical process of emancipation (Freire, 1970). The role of the educator—in this case, the AI-based conversational agent—is not to instruct, but to stimulate reflection and critical awareness. Interaction with the LLM should therefore promote critical digital and media literacy, enabling young people to analyse discourses, identify their biases, and construct their own counter-narratives.

Authors such as Giroux (2011) and Buckingham (2007) highlight the need to educate subjects who are capable of critically interpreting media messages and actively participating in the digital public sphere. From this perspective, the educational chatbot should be conceived as a space for dialogical and reflective learning, where language is used not only to inform, but also to foster empathy, deliberation, and resistance to intolerance, deliberate deception, and especially to the kind of malicious information found at the intersection of disinformation and hate speech.

Contemporary approaches to critical media literacy (Livingstone, 2014) emphasise the development of interpretative, ethical, and emotional competencies when dealing with potentially harmful content. In the context of the project, these competencies are linked to the ability to recognise, contextualise, and respond to hate speech in digital environments, for which the tool designed to navigate networks of hate speech and disinformation is intended.

The phenomenon of online hate speech has been extensively analysed from the perspectives of communication, social psychology, and media studies. Benesch (2014) uses the term counter-speech to refer to discursive strategies that confront hate through rational arguments, humour, empathy, or education. These strategies aim to replace punitive censorship with a pedagogy of dialogue and communicative resilience.

Furthermore, recent research (Mendel & Sharvit, 2019; Cammaerts, 2021) highlights the significant potential of pre-bunking and cognitive inoculation. That is, exposing users to attenuated versions of problematic discourses to strengthen their critical and emotional capacity against manipulation. These strategies are not limited to the exposure and analysis of the discursive logics behind the spread of hate speech and disinformation through the social network analysis (SNA) tool, but can also inform the pedagogical logic of the LLM, so that the system's responses do not merely correct or sanction, but actively model forms of reasoning and empathy in the face of hate speech.

2.4 Integration of Sociotechnical, Ethical, and Pedagogical Dimensions

The theoretical framework presented here ultimately allows for the articulation of three key dimensions in the design of the application: 1) the strictly sociotechnical dimension (Latour, 1987; Callon, 1986; Suchman, 2007), which emphasizes the role of participatory design in sociotechnical systems, highlighting the social contextualization of use and the social mediation of its design and implementation; 2) the ethical dimension (Floridi et al., 2018; Brey, 2012; European Commission, 2024), which focuses on the need to provide clear and

transparent explainability, promote equity, accountability, and human oversight throughout the entire process—from conceptualization and prototyping to use in socio-educational environments; and 3) the pedagogical dimension (Freire, 1970; Giroux, 2011; Buckingham, 2007; Benesch, 2014), which gives ultimate meaning to this sociotechnical development by situating it within a framework of dialogical learning, critical thinking, empathy development, and the facilitation of counter-narratives.

This integrated approach allows empirical findings—derived from focus groups, user tests, expert interviews, and sessions with young people and educators—to be translated into well-founded sociotechnical requirements that address the real needs of users while complying with European ethical and regulatory standards. Ultimately, the goal is for the HATEDEMICS application to be conceived not merely as a learning tool, but above all as a space for cultural and democratic mediation, where technology actively contributes to the construction of a critical and tolerant digital citizenship.

3 METHODOLOGY: THE CO-CREATION LIFECYCLE

HATEDEMICS adopts a socio-technical and mixed-methods research design, grounded in the assumption that hate speech, disinformation and conspiracy narratives are not merely "content problems", but socio-political phenomena shaped by institutional settings, platform infrastructures, legal frameworks and user practices. Consequently, defining socio-technical requirements cannot rely exclusively on technical feasibility or abstract theoretical models. Instead, it requires a systematic and iterative integration of conceptual, empirical, participatory and technical evidence.

From the outset, WP2 was designed not as a linear needs assessment exercise, but as a reflexive methodological process, capable of capturing contested meanings, behavioural dynamics and ethical tensions associated with the use of AI in socially sensitive domains. This approach is particularly relevant in the context of the social sciences, where categories such as "hate", "harm" or "risk" are inherently normative, context-dependent and politically contested.

To operationalise this approach, the project follows an iterative workflow across deliverables and work packages, ensuring cumulative learning and traceability:

- **Conceptual mapping and desk research (D2.1)** to define the analytical problem space, identify key debates and establish a shared conceptual language.

- **Empirical validation (D2.2)** through qualitative methods (focus groups and semi-structured interviews) to capture stakeholder needs, tensions and interpretative frames.

- **Participatory refinement (WP4 co-creation)** to test assumptions, translate needs into usable workflows and validate educational applicability.

- **Pilot-based and UX evidence (WP4 Phase 1)** to observe real user interaction patterns, capture behavioural reactions, identify misconceptions and friction points, and gather ideas and suggestions of activities that could be potentially included in the second version of the platform.

- **Technical iteration (WP3**) to align socio-technical requirements with platform components and constraints, ensuring implementability without reducing the socio-political complexity of the phenomena under study.

Crucially, the methodology integrates a comparative politics lens. Rather than treating national contexts as anecdotal case descriptions, the project leverages comparison to identify: (i) cross-cutting patterns that can be generalised into common EU-level requirements; and (ii) context-specific drivers that justify modularity, configurability or differentiated guidance within the platform and educational toolkit.

3.1 Desk research (D2.1): mapping the problem space and building a shared conceptual language

The desk research conducted under D2.1 provided the analytical foundation for defining the socio-technical requirements, functioning not merely as a literature review but as a problem-structuring exercise informing all subsequent methodological stages. This phase fulfilled four key functions.

First, D2.1 clarified and delimited the conceptual landscape of hate speech, disinformation, misinformation and conspiracy theories, highlighting their frequent overlap in real-world digital environments. This finding challenged rigid categorical approaches and informed the decision to prioritise contextualised and pedagogically oriented explanations in system design.

Second, the desk research analysed the policy, legal and governance frameworks relevant to platform deployment, including EU-level debates and national regulatory differences on freedom of expression and hate speech. These insights shaped early assumptions regarding data use, transparency requirements and user guidance necessary to ensure responsible and compliant use across contexts.

Third, D2.1 mapped the state of the art in AI-based content analysis and counter-speech tools, identifying recurring limitations such as bias, opacity, dataset dependency and multilingual constraints. Particular attention was paid to points of tension between technical systems and social-science concepts, including intent, irony, power relations and contextual meaning.

Finally, the review examined educational approaches to media literacy, critical thinking and civic resilience, with a focus on participatory and co-creative pedagogies. This informed the alignment between platform functionalities and the WP4 Training Kits, ensuring that socio-technical requirements are not only operationally sound but also pedagogically grounded.

3.2 Socio-technical v1 empirical validation (D2.2): focus groups and interviews

D2.2 marked the transition from conceptual mapping to empirical grounding by relying primarily on qualitative methods, namely focus groups and semi-structured interviews. This approach was adopted to capture dimensions that cannot be adequately assessed through quantitative or purely technical analyses.

The empirical work focused on how stakeholders define hate speech and disinformation in practice—often diverging from legal or academic definitions—how different user groups perceive harm, risk and priority within their socio-political contexts, and how professional communities navigate trade-offs between freedom of expression, minority protection and platform accountability. Particular attention was also given to stakeholders' expectations, concerns and ethical reservations regarding AI-based tools, especially in relation to explainability, bias and responsibility.

The sampling strategy deliberately prioritised heterogeneity, involving NGOs and civil society organisations, journalists and fact-checkers, educators and trainers, young people, and policy-oriented practitioners. This diversity is methodologically essential, as socio-technical requirements must accommodate varying levels of digital literacy and familiarity with AI, distinct operational constraints such as time pressure or safeguarding obligations, and a wide range of use scenarios spanning professional monitoring, educational settings and civic engagement.

Both focus groups and interviews followed semi-structured guides designed to balance cross-country comparability—central to the project's comparative politics approach—with sensitivity to national and sector-specific contexts, including legal frameworks, media ecosystems and organisational cultures. This design enabled the systematic identification of recurring patterns and divergences, which directly informed the logic of requirement building in terms of needs, tensions and design implications .

Table 1: Participation to focus groups and interviews

Country	Date	Format	Type
Italy	05/07/24	Focus Group F2F	Academics and members of civil society related to the research or target audience **(4 participants)**
Italy	12/09/24	Focus Group Online	Academics, OSINT analysts, journalists and experts on the subject **(4 participants)**
Italy	19/07/24	Interview Online	Project manager, expert on extremism, hate speech and counter narratives
Italy	23/07/24	Interview Online	Youth worker, No Hate Speech movement, National network against hate speech
Italy	11/09/24	Interview Online	Expert, policy maker
Italy	16/09/24	Interview Online	Journalist, fact checker
Italy	17/09/24	Interview Online	Journalist, fact checker
Spain	15/07/24	Focus Group F2F	NGOs/CSOs **(4 participants)**
Spain	24/07/24	Focus Group Online	Journalists, fact checkers **(8 participants)**
Spain	17/07/24	Interview Online	Anti-discrimination law expert (**8 participants**)
Spain	11/09/24	Interview Online	Expert, journalist
Spain	20/09/24	Interview Online	Expert, sociologist
Poland	17/07/24	Interview Online	Lawyer working on hate crimes and methods to counter them
Poland	24/07/24	Interview Online	Educator on human rights, anti-discrimination, hate speech and cyberbullying
Poland	25/07/24	Interview Online	Lawyer, anti-discrimination law expert, LGBTQ+ rights activist, law/policymaker
Poland	17/07/24	Focus Group Online	NGOs/CSOs **(7 participants)**
Poland	24/07/24	Focus Group Hybrid	Journalists, fact checkers **(6 participants)**
Malta	17/07/24	Focus Group Online	NGOs/CSOs **(7 participants)**
Malta	06/08/24	Focus Group Hybrid	Journalists, fact checkers **(10 participants)**
Malta	04/07/24	Interview Online	Expert, academic
Malta	10/07/24	Interview Online	Expert, fact-checker
Malta	15/07/24	Interview Online	Expert, academic

A key methodological outcome of D2.2 was the explicit recognition of conceptual ambiguities and controversies as analytically meaningful data. Rather than treating ambiguity as a problem to be eliminated, the project treats it as a design constraint: when users disagree on

definitions or boundaries, socio-technical systems must support interpretation, contextualisation and pedagogical clarification.

The empirical work surfaced recurring controversies, including:

- Where legitimate political critique ends and hate speech begins, particularly in polarised contexts;
- How to interpret humour, irony, coded language and "dog-whistles";
- Whether risk indicators might be misread as definitive judgements or predictions;
- Fears of censorship, misuse or over-policing of vulnerable communities;
- Tensions between demands for automation and demands for human accountability.

These controversies are not peripheral. They directly inform requirements related to explainability, confidence signalling, ethical warnings and human-in-the-loop governance.

3.3 Capturing real user interaction patterns: co-creation, pilots and UX testing (WP4)

While D2.1 and D2.2 provided conceptual and perception-based evidence, WP4 co-creation activities, Phase 1 and 2 pilots and UX testing introduced a complementary methodological layer focused on the observation of real user interaction and behaviour. This step is essential, as stated user needs often diverge from actual usage patterns when users engage with complex dashboards, indicators or AI-generated suggestions.

User interaction was analysed through five core usability dimensions: ease of use during first interaction, efficiency in task performance, memorability of interface use, frequency of errors and usage barriers, and overall user satisfaction. Failure to meet these criteria risks user disengagement and abandonment.

Co-creation activities functioned as a methodological bridge between research and design, serving to validate the relevance of emerging requirements, test terminology, interface logic and educational sequencing, and identify misunderstandings at an early stage, before they translated into design shortcomings.

The co-creation activities involved:

- Preliminary face-to-face (F2F) workshop in Milan with the consortium members.
- Online Workshops on the development of Educational Paths to be used with the Hatedemics Platform. A total of four educational path national workshops were conducted, one in different consortium countries.
- Online Workshops dedicated to data collection activities for the training of the AI tools to be implemented in the Hatedemics Platform.

Table 2: Educational path national workshops

Workshop	Date	Number of Attendees
Poland	13.11.2024	10 (all external)
Italy	18.11.2024	13 (7 internal, 6 external)
Malta	13.11.2024	10 (all external)
Spain	21.11.2024	16 (1 internal, 15 external)

Table 3: Participation to the Data Collection Workshop

Workshop	Date	Number of Attendees
Poland	21/11/2024	12
Italy	22/11/2024	13
Malta	27/11/2024	11
Spain	29/11/2024	13

These activities were carried out to understand the Potential Applications of the Hatedemics Platform, in particular by engaging with experts in diverse educational settings, we explored how the Hatedemics platform can be effectively used in real-world educational environments. The insights gained from co-creation workshops generated practical ideas and strategies, critical in shaping the development of the Hatedemics toolkit. Besides, the insights gained from co-creation workshops generated practical ideas and strategies, critical in shaping the development of the Hatedemics toolkit. By engaging with relevant stakeholders (i.e. NGO operators and fact-checkers from the consortium) we further discussed how to efficiently collect training material for the AI tools that have been implemented in the Hatedemics Platform.

3.4 Phase 1 pilots and UX testing as behavioural observation

Phase 1 pilots, supported by structured UX testing protocols, generated evidence on cognitive load and interpretative friction, behavioural responses to scores and indicators (ranging from overreliance to distrust), persistent misconceptions about AI capabilities (e.g. attribution of intent or legality), and educational scenarios requiring additional scaffolding and safeguarding. This evidence is central to D2.3, as it enables requirements to be defined not only in terms of functionalities, but also in relation to expected user behaviour and risk mitigation.

The project explicitly applies a comparative politics approach to enhance methodological robustness. This approach operates along two complementary logics. First, when similar needs and tensions emerge across most-different national contexts, they can be translated into core platform requirements, such as explainability, transparency and human control. Second, when national differences prove meaningful—due to legal thresholds, media ecosystems, target groups or platform preferences—requirements are designed as configurable modules or guided pathways rather than uniform solutions. This comparative lens enables credible generalisation, justifies modularity and localisation, and ensures alignment with EU policy expectations. A total of four pilots were conducted in phase 1, with two/three workshops in every of the previously-mentioned countries.

Table 4: Participation to the Pilot Phase 1 Workshops

Country	Date		Number of Attendees	
	W1 (F2F)	W2 (Online)	W1 (F2F)	W2 (Online)
Poland	27/05/2025	2/07/2025	11	17
Italy	06/06/2025	06/06/2025	12	9
Malta	27/06/2025	22/07/2025	10	8
Spain	22/05/2025	11/06/2025	12	13

From the phase 2 of the piloting, eight workshops were conducted: two in Italy (with 98 members the online one with 10), another in Spain (23 participants) and another with 16 participants in Spain, and another in Malta. By the time this deliverable is being made, we do not dispose of data about the Maltese pilot. Poland: two online with 6 and 11 participants and another F2F with 15 participants

Analytic pathway and traceability

To ensure traceability and accountability, HATEDEMICS follows an explicit analytic pathway linking evidence collection (D2.1 desk research; D2.2 focus groups and interviews; WP4 co-creation, pilots and UX testing) to thematic coding and synthesis, identification of socio-technical tensions (e.g. automation versus accountability, risk visibility versus fear amplification), and translation into functional and non-functional requirements. These requirements are subsequently validated through user feedback and pilot evidence, with Phase 1 results informing Phase 2 design choices and evaluation indicators. This process ensures that requirements are grounded in triangulated, comparative and behaviourally informed evidence rather than assumptions.

A key methodological strength of HATEDEMICS is that it extends beyond needs assessment to the testing of early functionalities and observation of user engagement. Evidence was gathered on navigation and interpretation of dashboards, understanding of detection outputs and confidence signals, perceived usefulness and trust in risk estimation, interaction with AI-assisted counter-narratives, educational sequencing and toolkit integration, and user reactions to ethical warnings and explanatory prompts. This behavioural focus is critical in a domain where AI risks being misinterpreted as authoritative truths.

Conceptual ambiguities and operational definitions.

WP2 and WP4 activities revealed persistent ambiguity among stakeholders regarding concepts such as hate speech, disinformation, misinformation and conspiracy theories. Rather than imposing rigid definitions, the project adopts a pedagogical and operational approach, treating concepts as contextual tools rather than fixed taxonomies. Socio-technical requirements therefore emphasise accessibility for non-expert users, pedagogical explanation through examples and scenarios, explicit acknowledgement of overlap and grey zones, and consistency of terminology across platform components and the educational toolkit. Pilot evidence shows that unclear definitions lead either to over-reliance on AI outputs or outright rejection; consequently, requirements specify embedded explanatory layers, explicit links between definitions and platform outputs, and a shared conceptual vocabulary across WP2, WP3 and WP4.

Functional requirements, jointly developed by FBK, CENTRA and FUNDEA, balance technical feasibility with socio-political constraints and are validated through co-creation and pilots. Data collection and monitoring functionalities must be transparent, documented and limited in scope, prioritising pattern and narrative analysis over individual profiling and clearly communicating representativeness limits and biases. Detection and analytical components are designed as interpretative aids rather than authoritative judgements, avoiding binary classifications, providing confidence levels and contextual explanations, and allowing users to explore why content is flagged.

The HATEDEMICS Risk component is defined as a contextual and educational indicator rather than a predictive or normative assessment. Requirements emphasise explicit uncertainty

signalling, explainability of contributing variables and interactions, and adaptation of visualisation and explanatory depth to different user profiles. Educationally inspired, non-alarmist visual metaphors are prioritised to foster understanding without fear amplification.

Counter-narratives, human-in-the-loop and educational integration.
Counter-narrative generation is framed as AI-assisted support rather than automation. Socio-technical requirements specify that counter-narratives are suggestions that remain editable, rejectable and embedded within workflows preserving human agency and responsibility. Full alignment with the WP4 educational toolkit is required, ensuring consistent pedagogical framing and integration into guided learning sequences. Ethical safeguards explicitly exclude individual targeting, automated dissemination and real-time intervention without human validation.
Taken together, these socio-technical requirements ensure that the HATEDEMICS platform operates as a coherent system in which technological capabilities, human practices and educational objectives are mutually aligned. They operationalise WP2 insights while remaining flexible enough to accommodate cross-country variation and future adaptation, providing a foundation for non-functional requirements, trade-off management and evaluation indicators.

3.5 From Field Exploration to Technical Solutions

The findings presented in this section were translated into specifications for Version 2 of the platform and the toolkit. To this end, work focused in particular on the following pillars.

Addressing Socio-technical Tensions (Lessons from Phase 1)
Based on the lessons learned during this exploration and piloting phase, changes were introduced in three improvement areas related to the socio-technical tensions that emerged:

- **Mitigation of algorithmic distrust:** Users reported a “fear” that AI might respond on their behalf. Version 2 establishes AI as an assistive tool rather than an autonomous agent, always ensuring human control.
- **Linguistic and contextual adaptability:** Specific features identified in Italy, Spain, Poland, and Malta have been incorporated, ensuring that hate detection recognizes local slang or expressions that often bypass standard filters.
- **Reduction of moderation fatigue:** The methodology introduces filtering criteria to prevent user stress caused by handling large volumes of data.

Specific Requirements for Disinformation Management
Considering the convergence between hate speech and fake news, Version 2 has incorporated technical requirements to “break” the disinformation cycle:

- **Source traceability:** The system should prioritize identifying the underlying narrative (the “story” behind the hoax) rather than focusing only on isolated messages.
- **Integration of verified content:** The platform integrates content from fact-checking agencies.
- **Inoculation strategy (prebunking):** The methodology promotes the design of educational strategies that anticipate common disinformation tactics before they escalate into hate crimes.

Linking Findings and Platform Design from a Socio-technical Perspective

Accordingly, the design of Platform V2 has taken these dimensions into account when developing the solutions. The following Socio-technical Matrix summarizes the link between each identified need and its corresponding technical or pedagogical response in this new phase of the project.

Table 5: Needs and solutions in Hatedemics Methodology v2

Dimension	Identified Need (D2.2 / Pilots)	Design Requirement (Methodology)	Implementation in V2 (WP3/WP4)
Human Agency	Fear of inappropriate AI responses.	Human-in-the-loop (HITL).	Validation panel and manual editing of counter-narratives.
Explainability	Lack of trust in detection criteria.	Algorithmic transparency	Confidence scores and justification labels in the dashboard.
Disinformation	Difficulty in debunking local hoaxes.	Fact-checking integration	Integration of content from fact-checking agencies into the platform.
Moderation Exposure Risk	Potential stress due to exposure to severe hate content.	Severity filtering	Risk ranking system to prioritize critical cases.
Prevention	Need to stay one step ahead of hate speech.	Pedagogical prebunking.	Educational modules that anticipate toxic narrative structures.

4 TECHNICAL REQUIREMENTS FOR THE HATEDEMICS PLATFORM v2

The second version of the platform was the result of a close collaboration among all partners. It involved, on the one hand, improving the platform's first version; on the other hand, introducing the educational activities adapted from the educational toolkit. A detailed description of all the changes made from Platform V1 to V2 is provided in D3.2. Below, we will highlight in particular how the collaboration with the partners from different WPs was fundamental for developing V2.

4.1 Platform improvement with respect to V1

After the pilots and the collection of feedback in WP4, the FBK team responsible for platform development in WP3 reviewed the results and prioritised the issues most frequently reported. The resulting changes addressed multiple aspects of the platform. While all the platform's modifications are described in detail in D3.2, below we will focus especially on those requiring close collaboration with the other partners. These improvements were implemented at three levels: high-level changes affecting the overall platform; enhancements to the Telegram data exploration module; and improvements to the counterspeech writing module.

4.1.1 High-level platform improvements

Platform translation: Translating the platform from English into the four project languages (Italian, Spanish, Polish, and Maltese) required close coordination with all partners. The JSON file containing the English-language functionalities of the platform was converted by FBK into a shared Google Spreadsheet accessible to all partners. Automatic translations were generated for each language using the `=GOOGLETRANSLATE` function, with one column per language. Partners reviewed these translations against the English source and revised them where necessary. Once the review was completed, the platform was updated, and partners tested it to verify the correctness of the translations.

Target redefinition: Following discussions with partners, the original labels used to identify marginalised groups targeted by hate were revised to adopt more inclusive terminology. The table below presents the original labels alongside the updated terms selected for each language, based on input from partners in the respective countries.

Table 6: Labels and translations in the platform's working languages

ORIGINAL LABEL	FINAL LABEL				
EN	**EN**	**IT**	**MT**	**PL**	**ES**
Women	-	Donne	Nisa	Kobiety	Mujeres
Jews	-	Ebrei	Lhud	Żydzi	Hebreos

Migrants	-	Migranti	Migranti	Migranci	Migrantes
Muslims	-	Musulmani	Musulmani	Muzułmanie	Musulmanes
Disabled	People with disabilities	Persone con disabilità	Individwi b'diżabilità	Osoby z niepełnospra wnościami	Personas con discapacidad
LGBT+	LGBTQIA+	LGBTQIA+	LGBTIQ+	LGBTQIA+	LGBTQIA+
POC	Black people	Persone nere	Nies suwed	Czarnoskórzy	Personas negras
Roma	-	Rom	Roma	Romowie	Personas gitanas

4.1.2 Components improvement: Telegram Data Exploration

The development and improvement of the specific components related to Telegram data exploration was carried out by FBK through an iterative validation process. This included both the Telegram data collection and the automated analyses included in the platform, which required review from other partners.

Telegram data collection: The initial dataset was built starting from a set of Telegram channel seeds provided by project partners with specific expertise in misinformation, ensuring that data collection was grounded in field knowledge and reflected relevant thematic, linguistic and socio-political contexts. In particular, partners were asked to share at least one channel per language which could possibly contain hate and/or disinformation against one of the marginalised groups of interest for the project. This initial set of channels was later expanded with the snowballing technique, as described in D3.1. This partner-driven seeding phase played a crucial role in orienting the subsequent expansion of the network and in anchoring the analysis to meaningful environments.

Automated analyses: For what regards textual analyses, initial exploration highlighted a number of challenges linked to the specificity of Telegram communication, including short conversational formats, high contextual dependency, multilingual content and the presence of non-textual elements (e.g. emojis, images). These features limited the direct transferability of models trained on other platforms and required a platform-sensitive methodological adjustment. To address these issues, the FBK team implemented a set of targeted improvements:

- **Hate speech and check-worthiness detection models refinement**. In particular, due to the absence of available Telegram datasets for Polish, a new, platform-specific annotated dataset was created in this language as a result of the collaboration between FBK and NASK. In this context, partners from NASK systematically validated automated outputs to identify recurrent misclassifications, language-specific biases and context-related errors.
  This collaboration not only strengthened the quality of the platform components, but also led to a joint publication between NASK and FBK that systematises the

methodological and empirical insights gained through the Telegram data exploration process (see: https://aclanthology.org/2025.clicit-1.56.pdf).

- **Topic modeling quality enhancement** through iterative optimisation of preprocessing steps, including lemmatisation, stop-word handling and emoji management. These methodological adjustments helped mitigate potential loss of context and avoid misinterpretations of the Telegram data during educational and pilot activities.

- **Revision of target identification**, balancing keyword-based strategies with qualitative assessment to reduce false positives while preserving interpretability.

- **Maltese language**: Following recommendations from Maltese partners, who noted that Telegram is not widely used in Malta and that there is no Telegram data available in Maltese, and considering the bilingualism of Maltese users, FBK decided to keep the Telegram data for Maltese in English, while still translating the platform's functionalities into Maltese.

Regarding network analyses, new measures are introduced in V2 following pilots' feedback, such as centrality measures and community detection. Thematic channel labelling is also introduced to improve interpretability and user experience.

4.1.3 Components improvement: Counterspeech writing

The dialogue system for assisted counterspeech generation is based on a model trained on a dataset of multilingual dialogues. These dialogues are fictitious exchanges between a person spreading hate and misinformation and an operator responding with counterspeech that challenges both hateful and misleading statements, using a fact-checking article as background evidence. The main aspect requiring partner collaboration was the collection of the dialogue data used to train the system.

Data collection was organised by FBK in successive stages, starting with simpler tasks and gradually increasing complexity for annotators. Initially, two internal annotators hired by FBK produced dialogues in English. These dialogues were then automatically translated into the target languages and reviewed manually by partners working in their respective languages. The resulting data were used to train the model deployed in the V1 platform.

For V2, partners were asked to create original dialogues directly in their own languages, again using a fact-checking article in the same language as reference. Across all phases, dialogues were produced with different degrees of automation within a human–AI collaboration setting. Three strategies were adopted:

- **Manual**: annotators fully composed the dialogue themselves, using the fact-checking article as a reference.
- **Pre-compiled**: annotators were provided with an automatically generated dialogue, which they reviewed and post-edited as needed. In this case, the automatically extracted source text supporting each operator turn was also provided.

- **Interactive**: at each turn, multiple automatically generated suggestions were dynamically proposed. Annotators selected the most appropriate option and could modify it freely.

Given the length of the data collection process, a dedicated working group was established to ensure continuous coordination with annotators. The group met biweekly to discuss ongoing and upcoming tasks, objectives, and to collect feedback, questions, and suggestions. It comprised 21 members, with all partner organisations represented and at least one participant per country attending the regular meetings. After each meeting, a recap email was circulated by FBK to keep all members informed. Working group members were either directly involved in dialogue annotation or coordinating colleagues engaged in the annotation process.

Feedback from the annotation working group was essential both for high-level decisions and for refining the dialogue collection process itself. First, following input from fact-checking partners, fact-checking articles were selected exclusively from organisations adhering to the International Fact-Checking Network (IFCN) Code of Principles, which was adopted as a shared quality standard. Annotator feedback also contributed to clarifying annotation guidelines and shaping the overall direction of data collection. Key examples include:

- **Dialogue translation**: annotators noted that, without explicit information about the country in which a news event occurred, proper contextualisation was often impossible. As a result, annotators were encouraged to provide additional context for elements unfamiliar to their national setting, particularly when referring to foreign institutions or events. Translation also raised the issue of unspecified gender; annotators were therefore encouraged to use gender-neutral forms when possible, and otherwise to balance feminine and masculine pronouns. This is done in line with EU inclusivity standards and to mitigate potential gender biases.
- **Interactive strategy**: When the interactive strategy was used to collect data for the V1 platform, the model generating dynamic suggestions at each turn often failed to maintain the role of the hate speaker because of its internal safety guardrails. For V2 data collection, a model fine-tuned on the V1 data was employed. Fine-tuning mitigated this issue and partially overcame the safety constraints, although some limitations and role inconsistencies still remained.
- **Pre-compiled strategy**: this approach proved more effective than the interactive one for role consistency, although it offered less flexibility to annotators in shaping the dialogue flow.
- **Maltese language**: specific measures were adopted based on feedback from Maltese annotators. As Maltese is a very low-resource language and current language models do not yet generate high-quality text in it, an ad hoc linguist was primarily tasked with translating English dialogues into Maltese.

Finally, another change resulting from discussions with partners concerned the labels used for speakers in the dialogue system interface. Originally, speakers were identified as “Hater” and “Operator”. Since labeling a person as a “hater” risks stigmatisation, and to avoid a “good versus bad” framing, the focus was shifted from individuals to content. This change also emphasises that spreading hate does not require specific personal traits and can involve

anyone. Accordingly, the labels were updated to “Hateful and misleading content” and “Counterspeech”.

4.2 Educational activities

A central element of the platform’s second version was the introduction of the educational activities, consistent with the Educational Toolkit (D4.3). This alignment was essential for preparing the platform for the second pilot phase, which would take place in different settings with young people and professionals working with young people or with educational backgrounds. To merge the toolkit with the platform, a specialised working group was set up. In addition to representatives from the FBK team, each country contributed at least one member with experience in education, and, when feasible, every partner organisation was represented. In particular, Maldita and CEO played a leading role in designing the activities, while FBK ensured their feasibility for implementation. The group held several focused meetings to maintain ongoing dialogue and to collaboratively integrate the toolkit’s materials into the platform. These continued discussions were essential to ensure that the user experience effectively supported and conveyed the activity’s learning objectives.

Following consultations with the WP4 leads, six activities were selected for the platform. These activities had two main objectives:

1) Cover the key learning points outlined in the educational toolkit, i.e. the activities named *Debunking*, *Lateral reading*, *Why We Care*, *Biases and Fallacies*.
2) Make the users familiar with the key concepts of the platform, i.e., the activities named *Spot the hate speech!* and *Spot the counterspeech!*

Two activities, *Debunking* and *Lateral Reading*, address misinformation, while the others focus on countering hateful content (activities such as biases and fallacies, factors contributing to the spread of hate speech and disinformation).

Every activity includes a theoretical overview and a hands-on component. The explanatory sections were either adapted from the Toolkit or created jointly with the educators. For the activity design and content, we adopted two different strategies. The hate-countering activities use identical material across all languages (English, Italian, Spanish, Polish, and Maltese) because the content is broad enough to fit each national context. In contrast, the misinformation-related activities required language-specific examples due to their reliance on contextualised fact-checking sources. The only exception is Maltese, which uses translated English examples because no Maltese fact-checking outlets exist, and the population’s bilingualism allows for this workaround.

Below, we provide a more detailed description of each included activity.

4.2.1 Debunking

The objective of this activity is to help users learn, in a guided way, some of the debunking techniques commonly used by professional fact-checkers.

The activity first presents a disinformation scenario, either an image with text contextualising the picture, and then guides the user through a series of closed-ended questions.

The questions included in the exercise are:

- *How would you verify?*
  This question aims to make the users familiar with the most widely used debunking techniques, such as reverse image search, advanced Twitter search, and AI-based tools. After the user answers and the feedback is given, they are shown the outcome of applying the correct technique (for example, the result of a reverse image search, including the image found).

- *What can you conclude by observing the result of the debunking?*
  This question is designed to encourage critical thinking and help users practice drawing conclusions based on evidence.

- *What kind of disinformation narrative can this hoax strengthen?*
  This question aims to familiarize users with common hoaxes targeting marginalized groups and to highlight how misinformation is often connected to hate narratives, even when this link is not explicit.

Each question is followed by feedback explaining why the selected answer is correct or incorrect. At the end of the activity, a final feedback message is displayed, linking to existing fact-checking articles that have already debunked the same piece of misinformation.

This activity was specifically designed for the platform based on input from Maldita. The specific disinformation scenarios used for each language were selected by the members of the working group from the respective countries. All scenarios are based on fake news that has already been fact-checked by a trusted organization and is linked to hoaxes targeting one of the marginalized groups addressed by the project.

4.2.2 Lateral reading

This activity is an adaptation of the lateral reading exercise described in D4.3, but modified to be done entirely on the Hatedemics platform. It is designed to teach transparency, source credibility, and how to identify trustworthy information.

Users examine screenshots of a website and answer questions about authorship, content and methodology, website security, and whether the content appears in other reliable sources. Multiple screenshots or a “Not found” option can be selected for each question, and users receive feedback explaining the correct answers. At the end, the activity provides an overall assessment of the website’s reliability based on the findings.

While the concept of the exercise was developed by educators for the educational toolkit, the activity was enriched with country-specific examples selected by members of the working group from the respective countries. The decision to include the activity on the platform was taken by Maldita’s team, while its design was developed collaboratively by the members of

the working group. Particular attention was also paid to designing clear feedback messages shown after each question, explaining why the selected screenshots were correct or incorrect.

4.2.3 Why we care and Biases and Fallacies

The objective of the *Why we care* activity (in the platform is labelled as Human Rights Declaration) is to develop knowledge about human rights and their connection to hate speech and misinformation. This is a group activity which allows participants to go through a selection of the human rights from the United Nations in the Universal Declaration of Human Rights (https://www.un.org/en/about-us/universal-declaration-of-human-rights), and to discuss how they relate to them.

Biases and Fallacies instead aims at raising awareness about social mechanisms, cognitive biases and logical fallacies and developing critical thinking competences. It consists of matching specific real life situations to the relevant logical fallacy or bias definition.

Both these activities closely mirror their homonymous activities described in D4.3 and developed for the educational toolkit by educators. No particular modification was made to any of them with respect to the original version presented in the toolkit.

4.2.4 Spot the hate speech!

The activity presents a series of messages (the same ones used in the toolkit), and users are asked to decide whether each message contains hate speech. As in the original toolkit activity, at the end of the activity users are asked how they would react to the different messages, followed by general feedback on the exercise.

While the concept of this activity follows the *Spot the hate speech!* exercise developed for the educational toolkit, it was adapted for integration into the platform. In addition to the toolkit version, this activity includes immediate feedback for each answer, explaining whether the user's choice is correct and why. These were developed in collaboration with the educators, who suggested including a similar activity.

The activity has two main objectives. First, it aims to help users distinguish between actual hate speech and expressions that may be offensive or controversial but do not meet the definition of hate speech, and to practice appropriate responses to hate speech. Second, the activity (Figure 1) is designed as a simplified version of the message table on the *Inside the Chats* page (https://hatedemics.smartcommunitylab.it/dashboard/discussion, Figure 2), allowing users to become familiar with the platform's navigation and the concepts it represents.

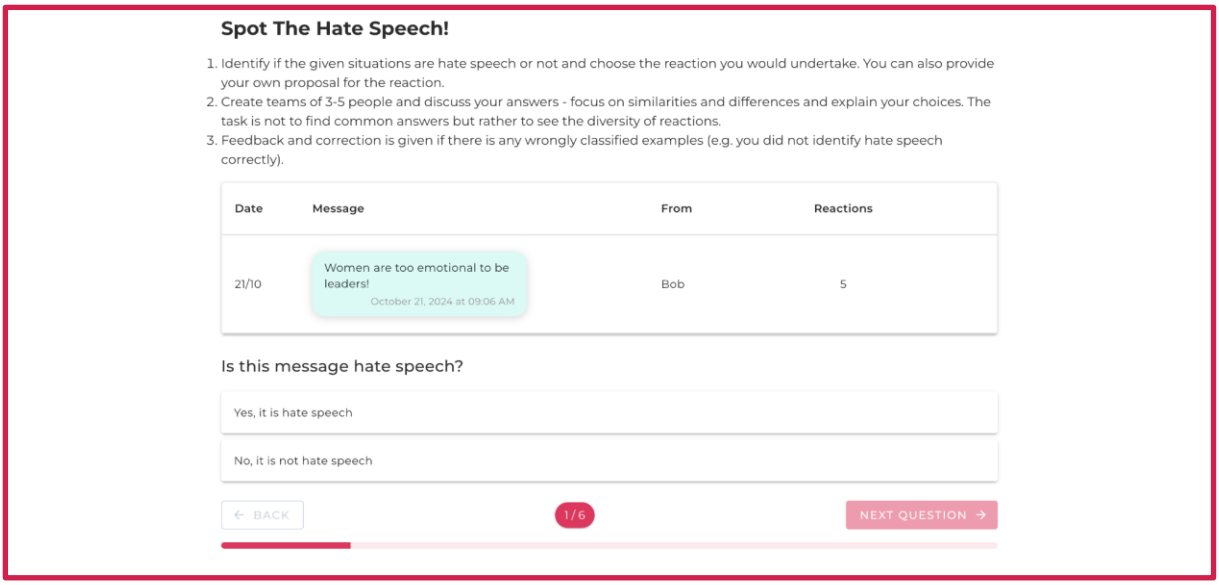

Figure 1: The Spot the Hate Speech! activity's page

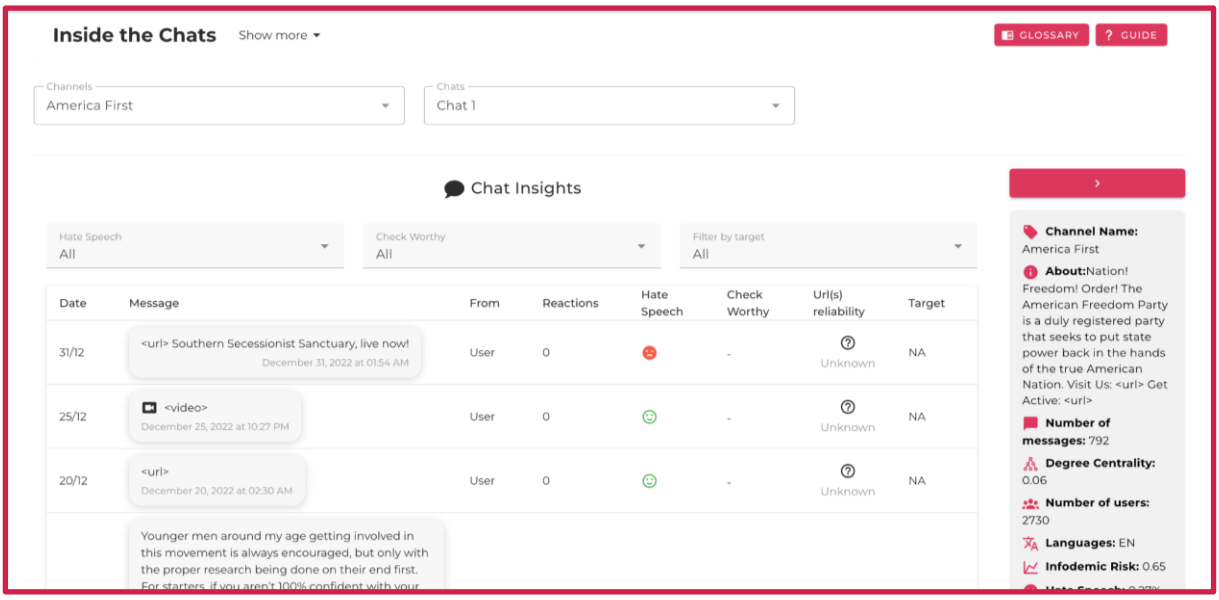

Figure 2: The Inside the Chats page

4.2.5 Spot the counterspeech!

This activity was specifically designed for the platform. The theoretical introduction presents the definition of counterspeech against hate and misinformation and outlines the key principles for effective counterspeech, such as avoiding abusive or divisive language, challenging the claim rather than the person, and using accurate, contextualised information.

The practical part of the activity presents a series of hateful or misleading statements, each paired with two possible replies. Users are asked to identify which reply represents appropriate counterspeech. After each selection, users receive feedback explaining why their answer is correct or incorrect. At the end of the activity, a summary of correct answers is displayed, along with recap questions similar to those used in the *Spot the hate speech* exercise.

The objective of the activity is to help users distinguish between genuine counterspeech and responses that, while opposing hate or misinformation, do not meet the criteria for counterspeech. It also allows users to practice appropriate reactions based on the guidelines introduced in the theoretical section. Additionally, similarly to the *Spot the hate speech* exercise, the objective is also to make users familiar with the concept of counterspeech, that they will need to properly understand other pages in the Hatedemics platform.

After developing a first draft of the examples to be included in the exercise, feedback from the working group highlighted the importance of referencing data used in counterspeech. As a result, sources were added to most counterspeech examples that include factual claims. The only exception is Example 2, where sources were intentionally omitted to avoid making the correct answer too obvious. In this case, the focus is on recognising divisive language and lack of empathy in one of the reply options, rather than on source verification. Below are two examples: one where sources were included and one where they were not.

Example 1 (including sources)
User: The world would be a better place without Muslims. They are only killing and raping our children.

Reply (counterspeech): On the contrary, most child abuse is operated by people they know: a relative, family friends, sports coach, someone in a position of trust and authority (source: Know Violence in Childhood, 2017).

Reply (not counterspeech): You are truly one stupid backwards thinking idiot to comment on Muslims like that.

Example 2 (not including sources)
User: Poland has the lowest number of terrorist attacks because they have a strict no-migrants policy. Migrants only bring trouble!

Reply (counterspeech): It's a misconception that Poland's low number of recorded terrorist attacks is solely due to a strict no-migrants policy since the premise inaccurately overlooks incidents that have occurred there since 2016, including attacks with anti-migrant motives. Moreover, research shows that migrants are less likely to commit crimes than native populations, and often, they themselves are victims of violence, demonstrating that the narrative linking migrants to terrorism is deeply misleading and harmful.

Reply (not counterspeech): I get that migrants can be irritating, but you're really exaggerating now! That's such a racist and backward way of thinking. Honestly, you should consider talking to a therapist, if a few smelly people make you this nervous, you're clearly oversensitive about the topic, and it shows.

5 EMPIRICAL FINDING AND SOCIO-TECHNICAL ANALYSIS

5.1 Key user needs and cross-country patterns

Building on the empirical evidence generated in WP2 (D2.2), co-creation activities and Phase 1 pilots and Phase 2 (WP4), it is observed how different user types interact with the platform, which needs emerge across national contexts, and how these needs translate into socio-technical tensions and, ultimately, into functional and non-functional requirements.

Rather than treating national cases as isolated examples, the analysis adopts a comparative EU-level approach, identifying recurring patterns as well as context-specific variations that inform the modular and configurable design of the HATEDEMICS platform.

Across the different national contexts involved in the project, several shared user needs consistently emerged, regardless of institutional setting or professional background:

- The need for explainable and transparent AI outputs, particularly regarding detection and risk indicators.
- The demand for contextualisation rather than binary classification, reflecting the political and social sensitivity of hate speech and disinformation.
- The expectation that AI tools should support, not replace, human judgement.
- The importance of educational framing, especially when tools are used beyond expert communities.
- Concerns about ethical misuse, including surveillance, censorship or reinforcement of biases.

At the same time, the comparative analysis also revealed context-specific variations that are methodologically and practically relevant:

- Differences in legal thresholds and public discourse norms affect how users interpret hate speech classifications.
- Variations in media ecosystems and use of the platform shape expectations regarding monitoring and data sources.
- National differences in institutional capacity and digital literacy influence how much guidance and scaffolding users require.

From a comparative perspective, these findings justify a hybrid design approach: a common core of requirements applicable across EU contexts, combined with modular features, configurable parameters and differentiated user guidance.

5.2 User workflows and interaction patterns

Empirical evidence from focus groups, co-creation workshops and pilots allows the reconstruction of typical user workflows, which are essential for translating needs into design requirements.

Across user types, workflows generally follow a sequence that includes:

1. **Exploration and orientation**
   Users initially explore data, dashboards or examples to understand the scope and relevance of the content. At this stage, clarity of interface, definitions and onboarding support are crucial.
2. **Interpretation and sense-making**
   Users attempt to interpret detection outputs, network visualisations or risk indicators. This stage is particularly sensitive to misconceptions, overconfidence in numerical scores, or confusion regarding AI limitations.
3. **Action or reflection**
   Depending on the user type, this may involve drafting counter-narratives, preparing educational activities, informing reports, or facilitating discussion.
4. **Evaluation and feedback**
   Users reflect on the usefulness, credibility and ethical implications of the outputs, shaping trust in the system.

These workflows are not linear and may vary by context, but they highlight critical interaction points where socio-technical tensions emerge.

5.3 Socio-technical tensions

The comparative analysis reveals that user needs often translate into structural socio-technical tensions, rather than straightforward feature requests. Key tensions identified across countries include:

- **Automation vs human accountability**
  Users value efficiency but reject fully automated judgements in politically sensitive domains.
- **Analytical depth vs cognitive load**
  Detailed analysis increases insight but risks overwhelming non-expert users.
- **Risk visibility vs fear amplification**
  Highlighting risk supports awareness but may unintentionally amplify anxiety or stigma.
- **Standardisation vs contextual sensitivity**
  Common EU-level tools must still account for national and local specificities.

Recognising these tensions is methodologically crucial: requirements must aim not to "resolve" them, but to manage and mitigate them through design.

5.4 Milestone workflow

Based on the evidence above, the project follows a structured translation logic:

User needs → Socio-technical tensions → Design decisions → Requirements.

Examples include:

- The need for interpretability leads to requirements for confidence levels, explanatory prompts and visual cues in detection and risk components.

- Ethical concerns translate into human-in-the-loop governance, warnings and methodological guidance embedded in the platform.
- Educational needs justify guided workflows, modular complexity and toolkit integration.
- Cross-country variation supports configurability and adaptable user pathways, rather than rigid system behaviour.

These translated requirements are further specified in subsequent sections, distinguishing between functional requirements (what the platform does) and non-functional requirements (how it behaves, under which constraints, and with which safeguards).

6 ETHICAL, LEGAL AND APPLICABILITY CONSIDERATIONS

The HATEDEMICS platform operates at a critical juncture where social science and data engineering meet, necessitating a design philosophy that prioritizes human agency over algorithmic authority. A fundamental insight from stakeholder engagement is that the use of AI in analysing hate speech and disinformation is never value-neutral. Because AI tools in sensitive domains carry a form of symbolic authority, they risk unintentionally legitimizing specific interpretations while marginalizing others. To counter this, the platform's socio-technical architecture is built on the principle of reflexivity, constantly reminding users that outputs are analytical aids rather than objective truths. This is supported by a commitment to transparency regarding system limitations, particularly what AI cannot infer, such as intent or moral responsibility. Ultimately, the responsibility for interpretation and action remains with the user, a design choice that pilot results show increases ethical trust by acknowledging uncertainty rather than claiming complete neutrality.

Given its strong educational orientation, HATEDEMICS treats learning contexts with specific care, acknowledging that power asymmetries and interpretive vulnerabilities are more pronounced in these settings. Ethical requirements specify that educational use must be framed as critical inquiry rather than a search for "correct" answers. Consequently, analytical outputs are designed to be accompanied by discussion prompts and methodological guidance for trainers, ensuring that AI-supported analysis enhances critical thinking rather than reinforcing authority bias. This pedagogical approach extends to safeguards for youth audiences, where age-appropriate language and visual separation between analytical exploration and normative judgment are mandatory. In such contexts, supervised use and guidance for managing emotional or polarizing content are essential to ensure that interaction with sensitive material remains a learning experience rather than a source of harm.

Addressing bias is perhaps the most prominent ethical concern within the stakeholder community. There is a persistent apprehension that AI systems might reproduce stereotypes, disproportionately flag marginalized communities, or reflect dataset imbalances. HATEDEMICS responds to this by adopting a bias-aware design philosophy rather than promising a bias-free system. This involves continuous monitoring, the documentation of known limitations, and a conscious shift away from individual-level profiling toward narrative-level and pattern-based analysis. By communicating potential biases and uncertainties clearly to the user, the platform treats bias mitigation as an ongoing and transparent process. This positioning is reinforced by explicit ethical warnings embedded within the user journey, which serve as pedagogical elements that discourage punitive or discriminatory use while clarifying appropriate versus inappropriate applications of the platform.

From a legal perspective, the platform ensures full compliance with GDPR principles, including lawfulness, purpose limitation, and data minimization. Compliance is treated as more than a backend technicality; it is a matter of transparency that allows users to understand exactly what data is processed and at what level of aggregation. Furthermore, although HATEDEMICS is not a commercial platform, it aligns with the transparency and accountability obligations of the Digital Services Act (DSA). This alignment is reflected in the clear documentation of methodological choices and system behaviour. Additionally, the platform is designed to respect cross-country legal disparities regarding hate speech and freedom of expression. By

avoiding the use of country-specific legal thresholds as default classifications and instead providing contextual legal references, the platform supports cross-country applicability and prevents legal misinterpretation.

Finally, usability and security are treated as socio-technical requirements that directly influence the ethical effectiveness of the tool. Poorly understood interfaces increase the risk of misuse or cognitive overload. To mitigate this, the platform emphasizes progressive disclosure of complexity and guided onboarding pathways. Design choices include using visual cues that prioritize educational clarity over alarmism and encouraging reflective pauses during the interpretation of complex indicators. Security measures, such as role-based access control and secure authentication, are also viewed through an ethical lens, ensuring that the system provides responsible access without creating unwanted surveillance dynamics. Together, these elements ensure that HATEDEMICS serves as a robust, responsible, and transferable tool for addressing toxic online content.

7 RECOMMENDATIONS

This section addresses a twofold aim: on the one hand, recommendations to be taken into account on sociotechnical requirements when using an AI platform are made; on the other hand, recommendations on how to effectively fight disinformation and hate speech are compiled.

7.1 Sociotechnical requirements

Sociotechnical analysis allows us to conclude that the development of a language model–based application to counter hate speech among young people cannot be reduced to a purely technological or educational issue. The research demonstrates that the effectiveness, legitimacy, and sustainability of such a tool depend on the balanced integration of social, technical, and pedagogical dimensions, as well as the active participation of young people throughout all phases of the process.

The findings show that generative AI can play a significant role in promoting critical thinking, empathy, and digital literacy, provided it is designed from a relational and pedagogical ethic that prioritises guidance, reflection, and respect for cultural diversity.

In this regard, the sociotechnical approach adopted in the project offers a reference model that can be applied to other educational AI developments, based on three fundamental principles: transparency, participation, and shared responsibility.

7.1.1 Main Conclusions

1. **Sociotechnical design enhances the social legitimacy of AI.**

   The co-creation processes (Sanders & Stappers, 2008) and participatory testing (Spinuzzi, 2005) carried out in the HATEDEMICS project not only provide data on usability and effectiveness, but also generate a sense of collective ownership that strengthens young people's and educators' trust in the tool. The legitimacy of educational AI depends as much on its learning capacity as on its ability to "listen" to the social actors involved in the phenomenon in a broad sense.

2. **Human mediation remains indispensable.**

   Results show that young people perceive AI as useful when it complements—not replaces—the role of teachers, tutors, or social educators. The system's added value lies in its function as dialogical support: offering a space to explore, practice arguments, and learn to respond to hate speech under the guidance of human reference figures.

3. **Ethics and pedagogy are inseparable dimensions.**

   The system's behaviour cannot be considered ethically neutral: every response generated by the LLM conveys implicit values about what is acceptable, legitimate, or correct. Therefore, ethical principles such as non-maleficence, justice, autonomy, and explainability must be integrated into the pedagogical design, rather than treated as external or secondary requirements.

4. **Cultural diversity is a structural axis, not a variable.**

   Tests conducted in different countries highlight the need to adapt the chat's language, examples, and communicative codes to each national and educational context. Recognising cultural and linguistic nuances not only prevents bias but also strengthens the tool's inclusive and European character.

5. **The development of educational AI requires multi-level, multi-stakeholder governance.**

   The sustainability of the project requires coordinated involvement from institutional, technological, and educational actors. Governance should not be limited to ethical oversight but extend to the definition of public policies that integrate these systems into strategies for media literacy, digital citizenship, and extremism prevention

7.1.2 Strategic Recommendations

Based on these conclusions, the following recommendations are proposed to guide the deployment, evaluation, and sustainability of the system within the European framework:

a) Social and Ethical Recommendations

- Promote continuous youth participation in future design iterations through co-creation labs or communities of practice.
- Ensure communicative transparency of the system: explain to users, in accessible language, how the model works, what it can and cannot do, and how their data is protected.
- Establish a permanent ethics committee to oversee model updates, assess biases, and propose correction or improvement mechanisms.

b) Technical Recommendations

- Implement traceability and internal audit mechanisms for interactions, in compliance with European AI regulations (AI Act).
- Develop a continuous evaluation framework combining quantitative indicators (usability, satisfaction, errors) and qualitative indicators (reflectivity, trust, empathy).
- Adopt an open-source model or partial public documentation to promote transparency and scientific collaboration without compromising system security.

c) Pedagogical Recommendations

- Integrate the tool into structured educational programs linked to subjects such as citizenship, digital ethics, or media communication.
- Train teachers and social educators in the pedagogical use of AI through guides, workshops, or open resources explaining its potential and limitations.
- Foster scenario- and dialogue-based learning, using the chat as a catalyst to analyse real discourses, promote empathy, and practice constructive responses.
- Incorporate an emotional education module to help young people identify and manage emotions arising from exposure to hate speech.

d) Institutional and Sustainability Recommendations

- Align the tool with European policies on media literacy and combating disinformation, ensuring coherence with the objectives of the Digital Education Action Plan (European Commission, 2020).
- Explore partnerships with educational networks, NGOs, and digital platforms to expand the impact and replicability of the system across formal and non-formal environments.
- Ensure long-term updating and maintenance through collaborative governance models involving universities, public administrations, and technology centres.

7.2 Disinformation and hate speech

Section 5.2 comprehends a set of recommendations on additional strategies and good practices that stakeholders – NGOs/CSOs, public authorities, young activists, LEAs, journalists, media professionals and policy-makers – may adopt beyond implementing the proposed solutions by HATEDEMICS. In this sense, it aims to propose approaches and specific practices that can be effective in the fight against disinformation and hate speech.

As previously argued in other stages of the project, understanding how to prevent, counter, and react to hate speech and disinformation is paramount to create safer spaces for victims, build more equitable societies, and encourage societies to act. Thus, this section is divided in two: general recommendations for all stakeholders, and specific recommendations for a few sectors.

7.2.1 General recommendations for all stakeholders

General recommendations are structured following the six general principles of reaction against hate speech and disinformation: engage, educate, analyse, respond, support, and report.

1. Engage in existing networks

   Although there are several public, private and third sector institutions that have normally developed an extensive work in tackling and responding to hate speech and disinformation, several of them are still working in silos, impeding the coordination of common actions. ***Fostering multi-agency cooperation*** is therefore crucial for all stakeholders to improve community relations and interactions that lead to common understandings of the phenomena, build effective responses, and improve the procedures of reporting and investigating these crimes.

   In doing so, it is crucial to ***develop community engagement initiatives*** for enabling stakeholders to create partnerships among local organisations and building solidarity networks. Provided that there are very different sensitivities between the actors, trust should be promoted through initiatives such as intergroup dialogue.

2. Educate society about hate speech and disinformation

   Education has plenty of benefits among stakeholders as it contributes to identifying what is a hate crime, and profoundly understanding its effects and motivations. Educating

people on hate speech and disinformation can mainly be articulated in a twofold sense: promoting training, and raising awareness.

On the one hand, ***periodical training programs*** may help stakeholders to completely understand these phenomena, anticipate challenges and properly address them. In the case of the organisations that work with victims such as law enforcement agencies and third sector organisations, special emphasis should be made on incorporating victim support and trauma-informed knowledge that improve their interactions with victims. On the other hand, ***community-based awareness initiatives*** enhance the knowledge of the general population on the phenomena, fostering the solidarity of them with victims, and making reporting procedures visible for everyone. It therefore encourages proactive responses, and may reduce the impact of stereotypes and misinformation campaigns among citizens.

3. Analyse media messages and practise critical thinking

Developing critical thinking is not an end in itself, but a key competence that help people to prevent disinformation and hate speech by critically analysing the available data, context and circumstances in order to comprehend them in the best possible manner. This concept is thus interrelated to the development of digital skills, which is also known as ***media literacy***. By developing the capacity of distinguishing between true and false information, and of enhancing citizens' capacity of reflection, individuals can acquire a sense of responsibility and discernment that may enable them to discern among the overload of online information, and to better support hate crime victims.

4. Respond to hateful and disinformation posts

Responses to hateful and disinformation posts can be made through social networks and media following different strategies. ***Monitoring online hate speech*** contributes to gaining a better understanding of the phenomenon, drafting a more accurate picture of the state of the art, and designing tailored policies against hate speech and disinformation. Another accurate response may be the ***promotion of information verification platforms and specialised observatories*** that verify content and foster multi-stakeholder cooperation in this matter, empowering thus communities to tackle disinformation autonomously.

Disinformation and hate speech can be also addressed through the ***creation and promotion of counter/alternative narratives***. These initiatives not only contribute to dismantling the fringe narratives that are disseminated to contaminate the public discussion, but also offer alternative frames and lenses to understand the context and interpret the events they are facing in society. In this sense, alternative framings should always be developed in accordance with human rights' values.

5. Support victims

Dealing with people who have experienced hate speech or have become victims of hate crimes is always a sensitive issue that requires professionals involved to know how to intervene. In this sense, ***adopting a victim-centred approach*** is crucial to provide an appropriate response as it focuses on the needs and perspectives of the victims, empowering them to participate actively in the process. Following this approach implies providing tailored support services that address the specific needs of victims, helping them cope with trauma and navigate the aftermath of hate crimes. In a nutshell, the key

core of it is to stand out the relevance of listening to victims, validating their experiences, and ensuring they feel supported throughout the legal and recovery processes.

Supporting victims also means to **train community members on how to safely intervene** when witnessing hate speech or hate crimes, **establish clear accessible channels for reporting** hate crimes and incidents, and **provide comprehensive support services** such as helpline and assistance services from public authorities and law enforcement agencies.

6. Report hate speech and disinformation

Adopting a proactive approach of reporting hate crime and disinformation is a fundamental pillar of digital citizenship. By escalating these issues directly to administrators or, in more severe cases, law enforcement, citizens move beyond passive consumption to active protection of the public discourse.

7.2.2 Specific recommendations

A few further recommendations can be remarked in the cases of policy-makers and media professional

Table 7: Specific recommendations to policy-makers and media professionals

POLICY-MAKERS	MEDIA PROFESSIONALS
- ***Legal framework and Policy Reform:*** Governments should enact robust hate crime legislation paired with transparent data collection to ensure perpetrator accountability and build community trust through informed policy-making. - ***Public Policy Lessons:*** Authorities must prioritise the development of bias-free AI, sustained digital literacy education, and accessible support infrastructures to holistically combat disinformation and protect vulnerable populations.	- ***Reporting Protocols:*** Establish clear channels for reporting hate speech and disinformation to administrators or law enforcement to ensure immediate intervention and accountability. - ***Legislative Strengthening:*** Advocate for comprehensive hate crime laws and transparent data collection to deter offenders and provide a factual basis for policy improvements. - ***Technological Ethics:*** Implement continuous oversight and inclusive design practices to identify and eliminate algorithmic biases within AI and digital platforms. - ***Educational Resilience:*** Invest in long-term digital literacy programs that equip citizens with the critical thinking skills necessary to navigate and neutralise online disinformation.

8 CONCLUSIONS

D2.3 presents the main conclusions and key findings of the HATEDEMICS project with regard to its socio-technical requirements. It represents a decisive transition from an initial emphasis on understanding the phenomena of hate speech and disinformation to the definition of concrete operational requirements for an AI-based system. By consolidating conceptual research, empirical evidence, and participatory feedback, we provide a coherent framework intended to inform the platform's design, its educational implementation, and its long-term sustainability.

We do not conceive the platform as a neutral technological artefact, but rather as a socio-technical system embedded within social practices and ethical norms. Our approach is grounded in a human-in-the-loop paradigm, whereby AI is designed to support and enhance human judgement rather than replace it. This framework is explicitly aligned with core European values, including transparency, explainability, human oversight, and the protection of fundamental rights. Consistent with a critical pedagogical perspective, we prioritise the promotion of media literacy, critical reflection, and civic engagement over automated enforcement or sanctioning mechanisms.

We identify a set of core functional requirements structured around detection, risk estimation, counter-narratives, and educational integration. In terms of detection, we frame AI outputs as interpretative aids rather than authoritative or binary classifications, recognising the contextual and contested nature of harmful content. Risk indicators are conceived as contextual and educational signals that clearly communicate uncertainty and avoid alarmist interpretations. With regard to counter-narratives, we envisage AI as supporting the generation of counter-speech while ensuring that human agency and editorial control remain central. At the same time, we position the platform as a learning environment, integrated with educational toolkits and incorporating age-appropriate safeguards.

Our research highlights several socio-technical tensions that must be addressed through careful design choices. Although users value efficiency, they express strong reservations about fully automated judgements in sensitive political and social domains, revealing a fundamental tension between automation and accountability. We also identify the need to balance analytical depth with cognitive load, ensuring that AI-generated insights are communicated in ways that remain accessible to non-expert users. Furthermore, the visibility of risk indicators must be managed so as to raise awareness without inadvertently amplifying fear, anxiety, or social stigma.

From an ethical and legal standpoint, we adopt a bias-aware approach rather than claiming to deliver a bias-free system. This involves documenting known limitations and focusing on broader content patterns rather than individual profiling practices. We ensure that the platform is designed in compliance with the General Data Protection Regulation and aligned with the transparency and accountability obligations set out in the EU AI Act and the Digital Services Act. Reflexivity is embedded within the system architecture through continuous reminders that AI outputs function as analytical supports rather than definitive judgements, thereby fostering user trust through the explicit acknowledgement of uncertainty.

Finally, we emphasise the indispensable role of human mediation, particularly by teachers, tutors, and educators, in enabling meaningful and responsible use of AI-based tools. We also stress the importance of cultural sensitivity, arguing for modular and configurable

requirements capable of accommodating diverse legal frameworks and linguistic contexts across Europe. Sustained participation through co-creation with young people and educators is identified as a key condition for social legitimacy, while long-term sustainability is understood to depend on multi-stakeholder governance and coordinated action between technological, institutional, and educational actors within broader digital citizenship strategies.

The project demonstrates that it is possible to develop AI systems oriented toward the common good, provided that their design incorporates democratic values, cultural diversity, and emancipatory pedagogical principles from the outset. The proposed application does not merely offer automated responses; it facilitates reflective and critical learning processes on how to recognize, understand, and counter hate speech. In this sense, it represents a significant contribution to the development of a mature, resilient, and empathetic European digital citizenship. The sociotechnical approach applied here—based on co-creation, design ethics, and critical pedagogy—can serve as a replicable model for future European initiatives seeking to integrate AI into educational contexts in a responsible and participatory manner.

REFERENCES

Benesch, S. (2014). *Countering dangerous speech: New ideas for genocide prevention*. Dangerous Speech Project.

Braun, V., & Clarke, V. (2006). *Using thematic analysis in psychology*. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brey, P. (2012). *Anticipatory ethics for emerging technologies*. NanoEthics, 6(1), 1–13. https://doi.org/10.1007/s11569-012-0141-7

Buckingham, D. (2007). *Beyond technology: Children's learning in the age of digital culture*. Polity Press.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. In J. Law (Ed.), *Power, action and belief: A new sociology of knowledge?* (pp. 196–233). Routledge.

Cammaerts, B. (2021). *Discourse against hate: Counter-narratives and civic engagement*. European Journal of Communication, 36(3), 229–246. https://doi.org/10.1177/0267323121991613

European Commission. (2020). *Digital Education Action Plan (2021–2027): Resetting education and training for the digital age*. Publications Office of the European Union.

European Commission. (2024). *Artificial Intelligence Act*. Official Journal of the European Union.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). *AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations*. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., & Cowls, J. (2019). *A unified framework of five principles for AI in society*. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Freire, P. (1970). *Pedagogía del oprimido*. Siglo XXI.

Giroux, H. A. (2011). *On critical pedagogy*. Bloomsbury Academic.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Livingstone, S. (2014). *Developing social media literacy: How children learn to interpret risky opportunities on social network sites*. Communications, 39(3), 283–303. https://doi.org/10.1515/commun-2014-0117

Mendel, T., & Sharvit, K. (2019). *Countering online hate speech through prebunking and inoculation strategies*. Journal of Social Issues, 75(3), 731–754. https://doi.org/10.1111/josi.12300

Sanders, E. B.-N., & Stappers, P. J. (2008). *Co-creation and the new landscapes of design*. CoDesign, 4(1), 5–18. https://doi.org/10.1080/15710880701875068

Spinuzzi, C. (2005). *The methodology of participatory design*. Technical Communication, 52(2), 163–174.

Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions* (2nd ed.). Cambridge University Press.

Winner, L. (1980). *Do artifacts have politics?* Daedalus, 109(1), 121–136.