



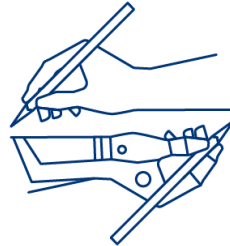
Funded by  
the European Union

Project: 101094364 — ITHACA — HORIZON-CL2-2021-DT-01018 (2021) 10433407 - 28/11/2025

EUROPEAN RESEARCH EXECUTIVE AGENCY (REA)

REA.C – Future Society

C.1 – Inclusive Society



**ITHACA**

AI To Enhance Civic Participation

**ITHACA**

**artificial Intelligence To enHance Civic pArticipation**

## D5.5: White Paper with policy recommendations

**Work Package:** WP5 – Conformity assessment tools policy recommendations and guidelines

**Authors:** SnP (Charikleia Eleni Nikolaou, Emmanouil Dimogerontakis), UniGraz (Jonas Seier, Maria Zangl, Michael Bedek)

RtF (Eva de Lera, Otilia Kocsis)

**Status:** Final

**Due Date:** 30/11/2025

**Version:** 1.0

**Submission Date:** 27/11/2025

**Dissemination Level:** PU – Public

### Disclaimer:

This document is issued within the frame and for the purpose of the ITHACA project. This project has received funding from the European Union's Horizon Europe Framework Programme under Grant Agreement No. 101094364. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

This document and its content are the property of the ITHACA Consortium. All rights relevant to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the ITHACA Consortium or the Partners detriment and are not to be disclosed externally without prior written consent from the ITHACA Partners. Each ITHACA Partner may use this document in conformity with the ITHACA Consortium Grant Agreement provisions.

(\* ) Dissemination level. - Public — fully open (automatically posted online)

Sensitive — limited under the conditions of the Grant Agreement

EU classified —RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444

## ITHACA Project Profile

Grant Agreement No.: 101094364

|                    |  |
|--------------------|--|
| <b>Acronym:</b>    | ITHACA   |
| <b>Title:</b>      | artificial Intelligence To enHance Civic pArticipation |
| <b>URL:</b>        | TBA  |
| <b>Start Date:</b> | 01/01/2023   |
| <b>Duration:</b>   | 36 months  |

### Partners

| Short Name | Legal Name  | Country |
|------------|---|---------|
| KT         | KONNEKT ABLE TECHNOLOGIES LIMITED                                 | IE      |
| CERTH      | ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS                | EL      |
| UPAT       | PANEPISTIMIO PATRON   | EL      |
| RtF        | RAISING THE FLOOR   | BE      |
| SnP        | STAMADIANOS KAI SYNETAIROI DIKIGORIKI ETAIREIA                    | EL      |
| UniGraz    | UNIVERSITAET GRAZ   | AT      |
| MNLT       | MNLT INNOVATIONS IKE  | EL      |
| SIMAVI     | SOFTWARE IMAGINATION & VISION SRL                                 | RO      |
| PEDAL      | PEDAL CONSULTING SRO  | SK      |
| BMA        | AGENTIA METROPOLITANA PENTRU DEZVOLTARE DURABILA BRASOV ASOCIATIA | RO      |
| MARTIN     | MESTO MARTIN  | SK      |



Funded by  
the European Union

## DOCUMENT HISTORY

| VERSION | DATE       | CHANGES  | RESPONSIBLE PARTNER                |
|---------|------------|----------|------------------------------------|
| 0.1     | 30/08/2025 | SnP      | First Structure/ToC                |
| 0.2     | 30/09/2025 | SnP      | Sections 1-4                       |
| 0.3     | 30/09/2025 | UniGraz  | Section 5                          |
| 0.4     | 15/10/2025 | SnP      | Section 6,8                        |
| 0.5     | 31/10/2025 | RtF      | Section 7                          |
| 0.6     | 05/11/2025 | All      | Internal Review                    |
| 0.7     | 18/11/2025 | KT, MNLT | Review                             |
| 0.8     | 25/11/2025 | SnP      | Final Version and Submission to KT |

# Table of Contents

- 1. Introduction..... 7
- 2. Reader’s Guide ..... 9
- 3. Problem Statement ..... 10
  - 3.1 High-level framing of the challenge ..... 10
  - 3.2 Democracy in the digital AI-driven era..... 10
  - 3.3 Role of CEPs..... 12
  - 3.4 Role of AI in CEPs ..... 12
- 4. Definitions & Taxonomy ..... 14
- 5. Core Values & Principles..... 18
  - 5.1 Participatory and Deliberative Democracy, Pluralism & Inclusion, Civic Engagement Platforms ..... 18
  - 5.2 Ethical Values ..... 19
- 6. Legal Landscape & Regulatory Context..... 22
  - 6.1 AI Act..... 22
  - 6.2 GDPR & ePrivacy..... 24
  - 6.3 Digital Services Act (DSA) ..... 25
  - 6.4 NIS2 & CRA ..... 26
  - 6.5 Accessibility frameworks: Web Accessibility Directive & European Accessibility Act 27
- 7. Insights from ITHACA..... 29
  - 7.1 Method in brief (tools, data, limitations) ..... 29
  - 7.2 Insights from workshops, focus groups and consultations..... 30
  - 7.3 Inclusion readiness vs accessibility ..... 31
  - 7.4 What didn’t work (and why) ..... 31
  - 7.5 Implications for policy recommendations ..... 32
- 8. Policy Recommendations for AI-enabled CEPs ..... 33
  - 8.1 Transparency, Accountability & Integrity ..... 33
  - 8.2 Privacy, Data Governance & Open Data ..... 36
  - 8.3 Responsible & Explainable AI for Deliberation ..... 38
  - 8.4 Inclusion, Accessibility & Deliberative Quality ..... 41
  - 8.5 Institutional Integration, Traceability & Platform Governance ..... 43
- 9. Conclusion ..... 48
- 10. References ..... 49

## ABBREVIATIONS

|               |   |
|---------------|---|
| <b>CEP</b>    | Civic Engagement Platform                   |
| <b>GDPR</b>   | General Data Protection Regulation          |
| <b>AI</b>     | Artificial Intelligence                     |
| <b>DSA</b>    | Digital Services Act                        |
| <b>DPIA</b>   | Data Protection Impact Assessment           |
| <b>CRA</b>    | Cyber Resilience Act                        |
| <b>FRIA</b>   | Fundamental Rights Impact Assessment        |
| <b>NIS2</b>   | Network and Information Systems Directive 2 |
| <b>VLOPs</b>  | Very Large Online Platforms                 |
| <b>VLOSEs</b> | Very Large Online Search Engines            |

## EXECUTIVE SUMMARY

In today's world, public debate and consultation occur through online platforms that expand participation while introducing new challenges to human rights, legitimacy, and institutional capacity. Civic Engagement Platforms provide traceable, and inclusive means of participation, when designed and guided by principles of transparency, fairness, and accessibility. Artificial Intelligence (AI) can enhance these platforms by assisting in their core functionalities.

This White Paper establishes a framework for the ethical, legal, and operational integration of AI in civic participation across the EU. It identifies key ethical values - transparency and explainability, accountability, fairness and non-discrimination, inclusion, and safety and security - and emphasizes three layers for effective AI-enabled participation: (i) process quality (visible tracking, timely feedback), (ii) responsible AI use (labelling, human oversight, explainability), and (iii) inclusion readiness (accessibility).

Its policy framework is organized around five pillars: (1) transparency and democratic accountability; (2) privacy, data governance and open data; (3) responsible and explainable AI for deliberation; (4) inclusion, accessibility and deliberative quality; (5) and institutional integration and platform governance. It situates these pillars within the evolving legal and regulatory landscape, including the AI Act, the General Data Protection Regulation, the Digital Services Act and accessibility EU directives.

# 1. Introduction

The present White Paper on Policy Recommendations (**White Paper**) for AI-Enabled Civic Engagement Platforms (**CEPs**) has been developed under the HORIZON-CL2-2022-DEMOCRACY project ITHACA (Grant Agreement No. 101094364). Purpose of this White Paper is to propose a coherent set of policy and governance principles that can guide the design, development, and operation of AI-driven civic technologies in ways that strengthen democratic participation, transparency, and institutional accountability.

Across Europe, civic engagement platforms increasingly rely on AI components to manage the scale and complexity of public input—supporting translation, summarisation, clustering, sentiment analysis, and moderation. These tools can help governments and communities navigate large volumes of citizen contributions, promoting more inclusive and informed decision-making. Yet, if not properly governed, they also raise substantial legal, ethical, and democratic risks, such as opacity in algorithmic decision-making, discriminatory outcomes, over-reliance on automation, digital exclusion, and limited institutional uptake of citizen feedback.

This White Paper responds to these challenges through a comprehensive policy framework that embeds ethical values and regulatory compliance within the technological and institutional design of CEPs. It provides actionable recommendations for legislators, public authorities, developers, and civil society actors to ensure that AI becomes an enabler of deliberation and inclusion—never a filter or substitute for democratic voice.

The policy recommendations are structured across five interrelated thematic units:

## 8.1 Transparency & Democratic Accountability.

CEPs must ensure that algorithmic processes and moderation systems are transparent, auditable, and explainable. Balancing algorithmic transparency with human oversight is essential to maintain civic trust. Platforms should disclose moderation rules, publish statements of reasons for content decisions, and enable meaningful appeals to reinforce democratic accountability.

## 8.2 Privacy, Data Governance & Open Data.

Privacy and data protection must be embedded by design and by default. CEPs should maintain clear GDPR-compliant governance through lawful basis mapping, data minimisation, and mandatory DPIA–FRIA assessments. Anonymised participation datasets should be responsibly released under the Data Governance Act and Open Data Directive, ensuring transparency, accountability, and public-interest re-use.

## 8.3 Responsible & Explainable AI for Deliberation.

AI used in civic deliberation must operate under robust accountability and explainability standards. CEP operators should maintain AI risk registers, ensure human-in-the-loop validation for automated decisions, and undergo regular external fairness audits. Explainable AI techniques should make algorithmic reasoning interpretable and contestable for both users and administrators, supporting informed participation and trust.

## 8.4 Inclusion, Accessibility & Deliberative Quality.

Inclusive participation and deliberative fairness are the foundation of democratic legitimacy. CEPs must comply with the Web Accessibility Directive and the European Accessibility Act, provide

multilingual and plain-language interfaces, and offer hybrid (online/offline) channels to reach all citizens—including persons with disabilities, older adults, migrants, and digitally excluded groups. Deliberative quality must be evaluated through fairness, representativeness, and institutional responsiveness rather than participation volume alone, ensuring that every voice counts and every contribution matters.

### 8.5 Institutional Integration & Platform Governance.

Civic participation should be systematically embedded in policymaking processes. CEPs must form part of formal consultation, planning, and budgeting cycles, with mechanisms for traceability and a “duty to respond” that links citizen input to policy outcomes. Transparent, rights-based procurement and open-source civic technology foundations will preserve public control, prevent vendor lock-in, and ensure long-term sustainability of democratic infrastructures.

The White Paper takes a rights-based and systemic approach that connects transparency, accountability, and explainability to the integrity of our democracies. The shared institutional responsibility of legislators, public authorities, developers, and civil society is at the core of the framework, each with a specific role to play in making sure that AI is an enabler of deliberation, and not a filter of our democracies.

The policy recommendations provide a normative foundation to the final ITHACA deliverable, the “CEPs Data Governance Framework”, which will materialize the principles in data governance rules and best practices. Together, they shape the ITHACA approach to creating trustworthy, lawful and participatory AI ecosystems in the civic space, showing how the digital transformation can be an instrument of democratic good in Europe by means of advancing transparency, fairness and civic empowerment.

## 2. Reader's Guide

### Target group

This White Paper targets decision-makers involved in shaping, funding, procuring, or operating civic engagement platforms across Europe. The main audience consists of EU policymakers and regulators, as well as national and local administrations and oversight bodies. This also pertains to civic-tech providers and integrators involved in design and deployment, along with civil-society and research organizations that monitor democratic quality, inclusion, and the protection of human rights. In organizations, this is particularly pertinent to policy owners, procurement officers, product leads, AI/IT architects, Data Protection Officers, AI Officers, as well as legal and governance teams.

### Goal

The objective is to convert the lessons learned from the ITHACA project into practical recommendations for enhancing AI-enabled civic participation. This approach links philosophical, ethical, and legal principles to design specifications, operational protections, and governance structures.

### Scope

In scope:

- A concise overview of democracy's challenges in the digital AI-driven era.
- Definitions and a role taxonomy aligned with EU law.
- A high-level overview of the digital regulatory landscape.
- A values-to-design baseline.
- Key insights from the ITHACA project.
- Thematic policy recommendations

### Out of scope:

Code-level specifications or vendor endorsements.

Templates for DPIAs, RoPAs, or Terms of Service, tailored to specific organizations.

Exhaustive legal commentary or case-law analysis.

### How to use this paper

Policymakers should consult the Introduction for key messages, and then the relevant thematic units in chapter 8.

Practitioners should start with chapters 4–6 to establish definitions, roles, and design controls, and then proceed to the recommendations in chapter 8.

## 3. Problem Statement

### 3.1 High-level framing of the challenge

Digital infrastructures are playing an increasingly significant role in shaping European democracies. Public debate, agenda-setting, and consultation now occur on online platforms characterized by limited attention, abundant information, and an increasing volume of input that administrations must manage. This fosters opportunities for increased participation and enhanced evidence, while also introducing new challenges to legitimacy, rights, and institutional capacity (OECD, 2023; European Commission, 2020).

CEPs facilitate more organized and verifiable methods of participation. When effectively designed, they gather proposals and feedback, facilitate volume management, and ensure that public engagement is traceable over time. In the absence of principled and law-abiding design, CEPs may exacerbate inequalities or lead to superficial interactions. It is essential to establish clear standards regarding transparency, inclusion, and feedback (Council of Europe, 2018; OECD, 2017).

AI has the potential to enhance these areas when implemented as an assistive layer rather than a replacement for public reasoning. The White Paper outlines that AI can summarize inputs, identify emerging themes, facilitate language translation, and assist in human content moderation. This also brings to light concerns regarding inclusivity, fairness, autonomy, as well as the manipulation of public opinion. In the European Union, AI assistive tools designed for civic participation should maintain transparency, be supervised by a human, and be subject to audit. This aligns with the AI Act, the Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law as well as other related legislative instruments, policy guidelines, and regulatory frameworks (European Parliament & Council, 2024; Council of Europe, 2024).

The White Paper translates philosophical, ethical and legal values into requirements for design, operational safeguards and governance framework, specifically for the enablement of participation by AI. It frames the challenges; aligns roles and concepts with EU law; overviews the legal landscape; and, in chapter 8, formulates feasible recommendations, in line with the Commission's better regulation agenda (European Commission, 2021, 2023).

### 3.2 Democracy in the digital AI-driven era

Digital technologies are reshaping the ways in which citizens interact with public institutions. The participatory potential is considerable; however, it also poses risks to trust, fairness, and legitimacy if individuals feel unheard or are unable to recognize the impact of their contributions. Genuine and deliberative participation is essential; otherwise, it risks becoming merely a symbolic exercise or a mere formality, which can lead to a decline in public confidence in democracy.

Challenges related to trust and legitimacy arise when participation is perceived as performative, unresponsive, or merely tokenistic. The input into digital processes lacks transparency; individuals are unable to observe how their contributions are managed, how competing inputs are analysed and aggregated, or the reasons behind the selection of certain ideas over others. In the context of the EU, the principles of timeliness, fairness, and transparency are grounded in the right to good administration as outlined in Article 41 of the EU Charter of Fundamental Rights, as well as in the

*Better Regulation Agenda* of the European Commission. Failing to "close the loop" on public participation leads to diminished engagement and an increase in skepticism.

The inequalities of participation in online spaces can mirror those present offline and may be exacerbated by a lack of inclusive design. Language and literacy, disability, connectivity, devices, and various economic or social barriers all contribute significantly to the issue at hand. The Web Accessibility Directive, the harmonised EN 301 549 standard, and the recent European Accessibility Act (EAA) establish essential accessibility criteria that public-sector digital services are required to fulfill under Union law. Compliance with these standards is necessary to prevent the systematic exclusion of individuals with disabilities and speakers of minority languages (European Parliament & Council, 2016; ETSI, 2021). For participation to be accessible and inclusive, it is essential to incorporate multilingual support, clear language, and assistive user interfaces from the outset, rather than as an afterthought.

Information disorder, manipulation, and coordinated inauthentic behavior can negatively affect the quality of deliberative participation and require effective management. Polarizing and toxic discourse, along with brigading campaigns, can overshadow more thoughtful perspectives and significantly influence visibility. The systemic risks associated with online platforms are increasingly acknowledged in EU policy, particularly through the due-diligence requirements and transparency measures outlined in Code of Conduct on Disinformation under the Digital Services Act. Safeguards should align with the principle of freedom of expression to prevent hindering debate or infringing upon rights.

The transparency of digital processes significantly influences perceptions of fairness. Algorithmic functions such as ranking, de-duplication, and automated summarization can enhance participation scalability. However, a lack of understanding regarding how these processes influence visibility and participation outcomes can lead to a decline in trust. The AI Act advocates for transparency, human oversight, and record-keeping as essential safeguards for AI-assisted functions and content moderation. This aligns with findings on governance, which indicate that effective, open, and accountable administrations are crucial for societal resilience. Such administrations enable governments to anticipate, absorb, respond to, and adapt to adverse events, thereby enhancing their crisis preparedness (European Parliament & Council, 2024; OECD, 2023). In the civic context, the explainability and contestability of these systems are essential prerequisites for their legitimacy.

Ultimately, the capacity limitations within public administrations represent a significant constraint. Public bodies operate within legally and logistically intricate environments, necessitating the management of substantial volumes of multilingual input. The responsibilities outlined intersect with various obligations related to cybersecurity as stipulated by NIS2, as well as procurement and accessibility requirements, and the *Better Regulation* expectations concerning evidence-based policymaking (European Parliament & Council, 2022; European Commission, 2021; OECD, 2023). The challenge may extend beyond merely increasing technology use; it involves ensuring alignment among legal obligations, democratic values, and operational capacity to prevent conflicting directions.

The strategic policy implication is straightforward: Europe should adopt digital participation to enhance representation and governance, while ensuring that human rights and legitimacy are not compromised. This White Paper asserts that AI and online platforms ought to serve as supportive infrastructure for democratic practices rather than as decision-makers. This aligns with the European Democracy Action Plan (EDAP) and the Commission's *Better Regulation* agenda, which emphasizes consultation, feedback, and evidence-based policymaking (European Commission, 2020, 2021, 2023).

### 3.3 Role of CEPs

CEPs serve as the vital link between citizens and administrative bodies. They offer a centralized platform for submitting proposals, comments, and endorsements; organize substantial amounts of input; and ensure a verifiable record of participation. European standards of participation and deliberation articulate these functions normatively, reflecting a broader dedication to transparent, inclusive, and accountable public decision-making (Council of Europe, 2018; OECD, 2020). Well-designed, CEPs can facilitate a transition from sporadic participation to ongoing, documented dialogue rather than a singular event.

**Reach** is the first dimension of value. Effective CEPs are characterized by low barriers to entry, achieved through mobile-first design, multilingual access, and straightforward onboarding processes. Accessibility-by-design serves as a fundamental legal and ethical standard for public entities within the EU, supported by the Web Accessibility Directive, the EN 301 549 standard (European Parliament & Council, 2016; ETSI, 2021), and the EAA. The focus is on ensuring equitable discoverability and fostering meaningful contributions within participating communities.

The second is **structure**. Unstructured input is transformed into administratively relevant input through tagging, clustering, de-duplication, and clear status transitions. Targeted analysis and more equitable comparisons between rival solutions are made possible by this approach. It also supports the design attributes outlined in international norms of deliberation, such as time to consider the evidence, access to fair information, encouragement of reason-giving through facilitation, and so forth (OECD, 2020). Scaling up without this framework soon becomes unfeasible for officials and participants alike.

The third is **traceability**. Participants should be able to trace their contribution via a clear procedure (received → considered → accepted/rejected → implemented) with timestamps and a short explanation at each stage. Because it implements the right to good administration, "closing the loop" is a popular phrase. Additionally, it recalls Better Regulation's openness guidelines for feedback and consultation (European Parliament & Council, 2012; European Commission, 2021, 2023). The opposite of the perception of performative engagement is traceability.

**Institutional fit** is the ultimate factor that determines success. Records management, security requirements (incident management, etc. under NIS2), and data protection measures (role-based access, minimization, legitimate basis under the GDPR, etc.) must all be connected to CEPs. If not, the platform damages confidence by creating expectations that the administration is unable to fulfill (European Parliament & Council, 2016, 2022).

### 3.4 Role of AI in CEPs

AI should be considered primarily to be an assistance layer spanning the participation lifecycle. AI can group and summarize inputs, spot new topics, and provide representative points of view throughout the sense-making stage, which helps communities and officials comprehend extensive and intricate consultations. When used appropriately, these technologies can support the 'Better Regulation' initiative of the Commission, which emphasizes transparent, high-quality, and evidence-based consultation (European Commission, 2021). Making public reasoning manageable at scale is the goal, not replacing it. AI has the potential to facilitate inclusivity and accessibility as well. For participants with a range of linguistic and literacy difficulties, assistive interfaces, machine translation,

and plain-language rewriting can reduce access barriers. These steps can also assist platforms in meeting international ethical guidelines for AI as well as EU-level accessibility criteria (European Parliament & Council, 2016; UNESCO, 2021). They are necessary for equal voice in a multilingual civic setting, not just politeness.

Additionally, AI can support the protection of civic engagement platforms' security and integrity. Human reviewers can identify toxic language, signs of manipulation, and coordinated inauthentic behavior with the aid of assistive moderation. However, the guiding principles in this case are support and proportionality once more: moderation help should not turn into a de facto automated enforcement, which would jeopardize the protection of rights under the AI Act and GDPR. Additionally, studies on AI moderation models have demonstrated that explainability and transparency are still lacking in practice and that, in the absence of human oversight, prejudice or over-blocking risks can compromise legitimacy (Zangl et al., 2025; Goyal et al., 2025).

Integrity is intrinsically linked to security. CEPs face growing risks at the intersection of information manipulation and cybersecurity, including adversarial attacks on moderation models and coordinated influence operations that take advantage of vulnerabilities in platform infrastructure. Researchers and policy organizations have observed that effective resilience against hybrid threats relies on the integration of cybersecurity measures from the outset, in conjunction with moderation controls (Achara & Chhabra, 2025; World Economic Forum, 2025).

These benefits, however, are accompanied by governance implications. The AI Act establishes essential principles of transparency, human oversight, robustness, and accountability that providers of AI systems are required to adhere to (European Parliament & Council, 2024; Council of Europe, 2024). In a civic context, these foundational principles manifest as practical controls and provisions for users.

It is essential to manage specific risks: summaries may frame issues in a way that marginalizes minority perspectives; clustering can obscure important nuances; toxicity detectors might misclassify culturally specific or reclaimed language; and manipulation signals can produce false positives during organic mobilization.

Ongoing assessment and the ability to challenge are essential. International guidance for assessing digital democracy and deliberation advocates for a systematic evaluation of deliberative processes and their associated tools, incorporating quality criteria and feedback mechanisms (OECD, 2021). CEPs ought to establish comparable evaluation baselines for their AI functionalities. OECD (2021).

Implementing these guardrails involves prioritizing explainability, logging, and redress as fundamental design controls rather than viewing them as mere compliance considerations. Examples include the publication of plain-language AI feature notices, the labeling of AI-assisted outputs, the maintenance of structured logs for moderation decisions, and the assurance that any data processing involving AI is proportionate, with a clear lawful basis, data minimization, and access controls in accordance with the GDPR, as well as adherence to record-keeping expectations outlined in the AI Act. These controls serve to ensure that AI assistance is transparent, subject to challenge, and capable of being audited.

## 4. Definitions & Taxonomy

This section sets a clear, shared vocabulary for AI-enabled civic-participation platforms. Where EU law already defines a term, that definition is used; where it does not, the White Paper provides a working definition for the purposes of this document.

### **AI System (legal definition)**

A **machine-based system** is designed to operate with varying levels of autonomy, possibly adapting after deployment. It infers from inputs how to generate outputs (predictions, content, recommendations, decisions) that influence physical or virtual environments.

*(AI Act, Regulation (EU) 2024/1689, Art. 3(1))*

### **Auditability**

**Auditability** is the ability to show compliance and decision-making through structured logs and documentation (e.g. moderation logs).

### **Civic Engagement Platform (CEP)**

A **Civic Engagement Platform (CEP)** is a digital system designed to enable individuals and groups to submit proposals, comments, and endorsements; support deliberation; and route inputs into administrative or decision-making workflows with appropriate logging and public feedback. CEPs are not limited to government use: they may be operated directly by public authorities, on their behalf, or by non-governmental and civil society organizations pursuing public-interest objectives (Farina, 2014; Martínez-Gil et al., 2025).

### **Digital Accessibility**

Public-sector digital services must be **perceivable, operable, understandable, and robust**, following the Web Accessibility Directive, the harmonised European standard EN 301 549 and the EAA.

*(Directive (EU) 2016/2102; ETSI EN 301 549)*

### **General-Purpose AI (GPAI) model (legal definition)**

A ‘**general-purpose AI model**’ means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks, regardless of the way the model is placed on the market, and that can be integrated into a variety of downstream systems or applications” *(AI Act, Regulation (EU) 2024/1689, Art. 3(63))*

The AI Act introduces obligations for GPAI providers under Article 53.

The European Commission complements the obligations under the AI Act with the **General-Purpose AI Code of Practice (2025)**, which details the manner in which providers of general-purpose AI models and of general-purpose AI models with systemic risk may comply with their obligations under the AI Act.

The ITHACA platform uses “GPAI” when referring to a component integrated into CEP functions (e.g., translation, summarisation).

*(AI Act; EC GPAI Code of Practice, 2025)*

### **Explainability**

The ability of an artificial intelligence (AI) system to make its decisions, operations, and results understandable to humans is known as explainability. It helps users and those who are impacted by an AI-driven decision, to understand how and why particular outputs are created, what circumstances led to them, and any potential restrictions or uncertainties.

It is not a standalone duty in EU law, but operationalized through obligations on transparency, user information, record-keeping, and human oversight.

*(AI Act safeguards)*

### **Toxic Speech Detection (TSD) Tools**

**Toxic Speech Detection (TSD) Tools** refer to AI-assisted classification systems that flag potentially abusive, harassing, or hateful content to aid human moderation. The output of a TSD tool is a decision-support signal, not an enforcement action in itself: final moderation must be made by a human, accompanied by notice and appeal rights.

### **Traceability**

**Traceability:** the ability to follow a contribution through all workflow states (received → reviewed → accepted/rejected → implemented), with timestamps and rationales. *(record keeping obligation under the AI Act; transparency reporting obligation under the DSA; accountability under GDPR)*

### **Participation Definitions**

For consistency across White Paper's chapters, the following key definitions are provided:

**Contribution:** any participant's input (proposal, comment, endorsement).

**Status state:** workflow label (received → under review → consolidated → accepted/rejected → implemented).

**Rationale:** a short plain-language explanation for a decision (e.g., why accepted/rejected).

**Moderation action:** decision applied to user-generated content or participant behavior on a digital platform

**AI-assisted output:** summaries, clusters, or flags generated by AI.

**AI-assisted feature:** CEP functionality powered by an AI system or GPAI (e.g., translation, summarisation, TSD AI Tool).

### **Role Taxonomy (aligned with EU Law)**

#### **Roles under the AI Act**

- **Provider:** means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge;
- **Deployer:** means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity;

- **Importer:** means a natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country;
- **Distributor:** means a natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market;
- **Authorised Representative:** means a natural or legal person located or established in the Union who has received and accepted a written mandate from a provider of an AI system or a general-purpose AI model to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation;

(AI Act, Art. 3)

### Roles under the GDPR

- **Controller:** means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;
- **Processor:** means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;
- **Joint Controllers:** when entities jointly determine purposes/means. (GDPR, Regulation (EU) 2016/679, Art. 4)

### Roles under the DSA

- **Intermediary services:** means one of the following information society services:
  - a 'mere conduit' service, consisting of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network;
  - a 'caching' service, consisting of the transmission in a communication network of information provided by a recipient of the service, involving the automatic, intermediate and temporary storage of that information, performed for the sole purpose of making more efficient the information's onward transmission to other recipients upon their request;
  - a 'hosting' service, consisting of the storage of information provided by, and at the request of, a recipient of the service;
- **Online platform:** means a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public, unless that activity is a minor and purely ancillary feature of another service or a minor functionality of the principal service and, for objective and technical reasons, cannot be used without that other service, and the integration of the feature or functionality into the other service is not a means to circumvent the applicability of this Regulation;

- **Very Large Online Platforms (VLOPs) / Very Large Online Search Engines (VLOSEs):**  
Services with  $\geq 45$  million average monthly active users in the EU, designated by the Commission, subject to enhanced systemic risk, transparency, and audit obligations.  
(*DSA, Regulation (EU) 2022/2065, Art. 3–45*)

## 5. Core Values & Principles

### 5.1 Participatory and Deliberative Democracy, Pluralism & Inclusion, Civic Engagement Platforms

CEPs extend the democratic public sphere online. Without value-aligned design and governance, toxicity and bias can silence under-represented voices and erode trust in deliberation (ADL, 2019; Falco & Kleinhans, 2018; Gil de Zúñiga, Jung, & Valenzuela, 2012). In CEPs, where inputs may inform policies, transparent, contestable moderation and bias-aware data practices are prerequisites for participatory and deliberative democracy, pluralism, inclusion, and civic trust. Explainable decisions and user-facing reasons increase rule comprehension and future constructive participation (Jhaver, Bruckman, & Gilbert, 2019; Molina & Sundar, 2022; Suzor, West, Quodling, & York, 2019).

#### **CEPs as one building block for participatory & deliberative Democracies:**

**Equal voice and reason-giving.** Toxic speech suppresses participation and narrows argument quality, effects that fall disproportionately on already under-represented groups (ADL, 2019; EISherief, Nilizadeh, Nguyen, Vigna, & Belding, 2018; EISherief et al., 2021). This undermines deliberative norms of inclusion and reciprocity that CEPs are meant to instantiate. Empirical work documents reductions in contribution and community exit following exposure to harassment, with particularly acute effects where identity-based hostility is involved (ADL, 2019; EISherief et al., 2018).

**From access to participation.** The OECD distinguishes information (one-way), consultation (two-way feedback), and engagement (collaboration across the policy cycle). CEPs should target genuine engagement: clear purpose, accountability for how input shapes decisions, transparency, and proactive inclusion measures (e.g., barrier reduction) to reach “silent majorities” and frequently excluded groups. These conditions improve representativeness and democratic learning (OECD, 2022).

**Balancing pre- and post-moderation.** For contentious, high-stakes debates typical of CEPs, a blended approach is recommended: pre-moderation (first-line shielding from harm) with timely post-moderation and accessible appeals to correct errors and preserve deliberative richness (De Gregorio, 2020). Public, human- or AI-generated explanations reduce recidivism and strengthen norm learning, which is critical for plural, reason-giving dialogue (Jhaver et al., 2019; Molina & Sundar, 2022).

**Countering cumulative inequities to ensure Pluralism & Inclusion.** Marginalized groups face structural underrepresentation and higher rates of targeted hostility online. CEPs therefore need zero-tolerance rules for identity-based abuse, paired with proactive design for accessibility and safety (ADL, 2019; Falco & Kleinhans, 2018).

**Human and algorithmic bias.** Moderation, human and automated, can misclassify dialects, reclaimed slurs, or identity terms, causing unequal false positive/negative rates. Inclusion requires bias-aware data practices and continuous evaluation. Explainable AI (XAI) methods that highlight toxic spans can make decisions contestable and corrigible, while audits track disparate impacts across protected attributes and languages (Halevy, Harris, Bruckman, Yang & Howard, 2021; Pavlopoulos, Sorensen, Laugier & Androutsopoulos, 2021; Ribeiro, Singh, & Guestrin, 2016; Waseem, Davidson, Warmsley, & Weber, 2017).

## 5.2 Ethical Values

This section sets out the ethical values that should be embedded in AI-enabled CEPs: transparency & explainability, accountability, fairness & non-discrimination (with inclusion), and safety & security. It builds on European and international guideline mappings (e.g., Fjeld, Achten, Hilligoss, Nagy & Srikumar, 2020; Jobin, Ienca, & Vayena, 2019) and aligns them with the EU’s emerging risk-based regulatory approach (AI Act, EAA) and data-protection principles relevant to high-risk deployments such as CEPs. In ITHACA, further steps have been made to operationalize such values, to make them auditable, measurable, and tied to concrete governance and technical practices, rather than treated as abstract aspirations (European High-Level Expert Group on AI [AI HLEG], 2019; European Commission, 2021; Fjeld et al., 2020; Jobin et al., 2019).

### Transparency and Explainability

Across widely cited guidelines, transparency and explainability include: open information about model logic and limitations; user notification when interacting with AI or when an AI makes (or informs) a decision about them; rights to information; and, where feasible, open data/code and regular reporting (Fjeld et al., 2020). These commitments appear alongside human oversight as core conditions for “trustworthy AI” (AI HLEG, 2019).

**Regulatory alignment for CEPs.** Under the EU’s risk-based approach, high-risk systems must provide information and transparency to users, together with documentation, logging, human oversight, and robustness/accuracy disclosures (European Commission, 2021). CEP deployments should therefore publish concise, comprehensible instructions for use, including intended purpose, expected accuracy/robustness, known failure modes, and conditions that may degrade performance (European Commission, 2021).

**Recommendations for ITHACA-like current and future initiatives.** A CEP should (i) provide user-facing reasons (“statements of reasons”) for removals/flags or automated assists; (ii) disclose model boundaries and known biases; (iii) log decisions for audit; and (iv) keep humans in the loop for contestation. The ITHACA project, in particular in the context of its Work Package (WP) 5 has already identified transparency and explainability as primary compliance priorities for conformity assessment, together with fairness, security, and privacy, so platform documentation and in-product explanations must be designed from the outset (AI HLEG, 2019).

**Participation, not opacity.** Given evidence that many AI ethics documents were developed in closed processes, especially in the private sector, ITHACA avoids “experts-only” transparency. Public-facing documentation and participatory oversight mechanisms (e.g., user panels) help alignment with inclusion and accountability goals (Schiff, Borenstein, Biddle & Laas, 2021).

### Accountability

Accountability in the guideline landscape spans verifiability/replicability, impact assessments, auditing and monitoring bodies, rights to appeal and remedy, and clear liability allocations (Fjeld et al., 2020). In practice, this means named responsibilities across the AI lifecycle and enforceable mechanisms, not just principles.

**Bridging the “principles-to-practice” gap:** Comparative reviews warn that ethics codes often lack real consequences: they are numerous, influential in discourse, yet frequently fail to shape real decisions or carry consequences for non-compliance (Hagendorff, 2020, 2022). ITHACA-like future CEPs might therefore build on regular compliance audits, publish periodic transparency reports, and define who does what (developer, deployer, independent reviewer) to make accountability actionable. (Hagendorff, 2020).

**Regulatory alignment for CEPs.** The AI Act requires risk management, data-quality controls, documentation, logging, transparency to users, human oversight, and robustness/accuracy for high-risk AI. These are directly applicable to CEP moderation/participation support and must be evidenced in conformity assessment (European Commission, 2021).

## Fairness and Non-discrimination

**What it means:** Guideline syntheses converge on i) preventing discriminatory outcomes; ii) ensuring representativeness, iii) high-quality data; and iv) building inclusiveness both in design and in impact assessment (Fjeld et al., 2020). Recurring risks identified in recent analyses, especially for GAI, are the amplification of societal bias, with under-representation and discrimination against minorities as likely harms if left unmanaged, and the intended spread of disinformation (e.g., Lockett, 2023; Abdelhalim, Anazodo, Gali & Robson, 2024; Weisz, He, Muller, Hoefer, Miles & Geyer, 2024; Dylan & Grossfeld, 2025).

**From principles to processes.** Reviews note that “operationalizable” principles (e.g., privacy, explainability, robustness) are often over-emphasized relative to harder issues (e.g., hidden labor, ecological costs); still, fairness and inclusion have increased in salience in the GAI era and must be treated as first-order requirements in civic contexts (Hagendorff, 2020, 2024). CEPs should therefore define fairness goals; test for disparate error rates across groups/languages; and demonstrate inclusive co-design and evaluation practices (Hagendorff, 2020, 2024).

**Regulatory alignment for CEPs:** The AI Act’s high-risk obligations require high-quality datasets and risk-mitigation to reduce discriminatory outcomes; combined with GDPR’s fairness and accountability ethos, this sets the baseline for data governance and model evaluation in ITHACA (European Commission, 2021).

**Inclusion in practice.** Evidence on how ethics frameworks are produced shows that public bodies engaged participatory processes far more than private actors. In line with the approach taken in ITHACA, similar future initiatives might institutionalize stakeholder participation, especially from groups most affected, when defining fairness criteria, test sets, and acceptable error trade-offs (Schiff et al., 2021; Akbarighatar, Pappas & Vassilakopoulou, 2023).

## Safety and Security

Technical robustness and safety cover secure-by-design development, reliability, predictability, and the prevention/mitigation of harm throughout the lifecycle (AI HLEG, 2019). Generative models however have the potential for risks: misinformation, deepfakes, adversarial attacks, leakage of sensitive data, and rapid shifts in attack modalities, therefore calling for continuous monitoring. (European Data Protection Supervisor [EDPS], 2024; Hagendorff, 2024; Weisz et al., 2024).

**Operational controls.** EDPS guidance recommends risk-based controls for GAI, staff training, and red-teaming to uncover unknown vulnerabilities, paired with ongoing monitoring and updates to risk assessments. Systems should also support fallback modes (e.g., switch to rules; hand-off to human-in-command) to maintain safety under adversarial conditions. (EDPS, 2024).

Implications for **ITHACA-like current and future initiatives.** “ITHACA’s WP5 integrated an AI cybersecurity tool to detect security breaches and hazards on Civic Engagement Platforms, grounded in qualitative and quantitative indicators, including behavioral and technical user activity data (‘user signals’), and big-data analytics. Such a tool should be paired with lifecycle risk management, incident response, and international collaboration between AI and security teams (AI HLEG, 2019).

## 6. Legal Landscape & Regulatory Context

This section details the EU regulatory frameworks relevant to AI-enriched CEPs, encompassing the AI Act, GDPR, E-Privacy Directive, DSA, NIS2, Cyber Resilience Act (CRA), and accessibility regulations, including the Web Accessibility Directive and the European Accessibility Act. This section offers a clear overview of each framework, outlining its definition and main objectives, the stakeholders involved, and its importance for CEPs and AI functionalities. This will also highlight several practical challenges to consider in the recommendations of chapter 8.

### 6.1 AI Act

#### What it is & why it matters.

The AI Act sets out a clear set of risk-based rules for AI developers and deployers regarding specific uses of AI. CEPs are anticipated to integrate diverse assistive AI elements to enhance user engagement, including translation and summarization features. Furthermore, these elements can assist in moderation and analysis by flagging toxic speech, supporting moderation through spam detection, clustering inputs, and conducting sentiment analysis. The AI Act creates a framework focused on risks that emphasizes transparency, human oversight, record-keeping, accuracy, robustness, and post-market monitoring obligations, shaping the design, labeling, and oversight of AI features. A municipality or Non-Governmental Organization (NGO) serves as a provider when it creates and promotes an AI system under its own name, and it acts as a deployer when it employs that system. Providers of GPAI models have distinct responsibilities, including the necessity for thorough model documentation, the maintenance of transparency, and the implementation of systemic-risk measures when necessary.

#### Risk-Based Approach

The AI Act is structured around a risk-based pyramid model. The responsibilities increase in relation to the potential risk to safety and fundamental rights.

**Prohibited AI systems:** includes practices such as social scoring, manipulative subliminal techniques, and the exploitation of vulnerabilities.

**High-risk AI systems:** (in the context of CEPs) are classified as such only when they meet the criteria outlined in Annex III of the AI Act or serve as a safety component as defined by the AI Act. In the context of democratic processes, a pertinent provision is point 8(b) of Annex III, which addresses *“AI systems intended to be used for influencing the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda. This does not include AI systems to the output of which natural persons are not directly exposed, such as tools used to organise, optimise or structure political campaigns from an administrative or logistical point of view.”*

**Limited-risk AI systems:** encompasses various functions pertinent to CEPs, including translation, summarization, topic clustering, and toxic speech detection. These systems necessitate transparency and user awareness initiatives, including the labeling of AI outputs, disclosure of user interactions with AI, and clarification of functional limitations.

In the context of **minimal or no-risk AI systems**, such as spell-checking and formatting suggestions, there are no binding obligations. However, the adoption of voluntary codes of conduct is recommended.

### **Examples of use cases and their classification under the AI Act**

- **Automated ranking or visibility scoring of contributions**

AI-driven ranking mechanisms are considered high-risk solely when they aim to affect electoral outcomes or voting behavior, as outlined in point 8(b) of Annex III of the AI Act. When ranking simply organizes or curates user submissions for discussion - without a clear intent to sway voting - the system is typically not deemed high-risk. However, it still falls under the transparency and accountability requirements of the Digital Services Act (DSA) for recommender systems.

- **Reputation / trust scoring of participants**

Systems that assign credibility and reliability to users pose significant risks when implemented in contexts outlined in Annex III, including access to essential services, education, employment, justice, or democratic participation. In such contexts, automated trust scores can significantly impact individuals' rights and opportunities. In different contexts, reputation indicators that are used solely for community moderation or to assess discussion quality are not considered high-risk according to the AI Act.

- **Targeted nudging to shape participation**

Analytics aimed at enhancing user engagement may be classified as high-risk if their objective is to sway voting intentions or electoral behavior, thereby establishing a direct connection between the system and democratic decision-making processes. Nudging, which aims to encourage or moderate civic participation, such as by suggesting that underrepresented groups increase their contributions, typically does not satisfy point 8 (b) of Annex III threshold.

### **Obligations for high-risk AI systems**

When a CEP is classified as a high-risk AI system, the provider must implement strong risk management practices, establish effective data governance, maintain comprehensive technical documentation, ensure human oversight, and guarantee the system's accuracy, robustness, and security. The provider must also perform a conformity assessment, register the system in the EU database, and monitor its performance after deployment.

Entities like municipalities and NGOs must use AI systems responsibly. When acting as public bodies or service providers, they are required to perform a Fundamental Rights Impact Assessment (FRIA) before their deployment. If significant changes are made to the AI system or if it is marketed under their own brand, they take on the responsibilities of the provider.

### **Practical guidelines for CEPs**

- It is important to make AI systems identifiable and understandable by notifying users when they interact with these systems and providing explanations in straightforward, accessible language.

- Maintain structured logs for traceability and accountability in AI-augmented moderation or content processing. These logs should include inputs, outputs, runtime parameters, and human interventions to facilitate post-market monitoring of performance, bias, and multilingual accuracy.
- Guarantee significant human supervision: Design CEPs that enable individuals to interpret, question, override, or halt AI outputs, particularly in decisions that impact their rights.
- Establish clear contractual AI roles and responsibilities: when incorporating GPAI or third-party AI modules, obtain documentation from providers and delineate the duties of providers versus deployers in contracts to ensure compliance throughout accountability chains.

## 6.2 GDPR & ePrivacy

### What it is & why it matters.

The GDPR sets the framework for processing personal data in CEPs. All data processing activities, including user registration, contribution logs, and moderation records, must adhere to GDPR principles such as lawfulness, fairness, transparency, purpose limitation, and data minimization (Voigt & Von dem Bussche, 2017). Controllers, usually public authorities, are required to establish a valid legal basis for these processing activities. They must implement privacy by design and by default, ensure secure processing, maintain a Record of Processing Activities (RoPA), and conduct a Data Protection Impact Assessment (DPIA) when processing is likely to pose a high risk to individuals' rights.

The ePrivacy Directive (2002/58/EC) safeguards the confidentiality of communications and establishes among other rules for the use of cookies and similar tracking technologies. In practical terms, non-essential tracking methods, such as analytics cookies, typically necessitate informed and freely-given consent according to EU law. Conversely, cookies deemed strictly necessary for the provision of services may be exempt from this requirement (European Parliament & Council, 2002; European Data Protection Board, 2020).

- **Data Protection Impact Assessment**

A Data Protection Impact Assessment (DPIA) is required where processing is likely to result in a high risk to the rights and freedoms of natural persons according to Article 35 GDPR. The risk level depends on the nature, scope, context, and purpose of the processing, mirroring the AI Act's risk-based logic that differentiates obligations according to potential impact.

For CEPs, a DPIA must be conducted when a data processing activity meets at least two of the following criteria:

- includes automated decision-making or profiling that may substantially influence participation or visibility;
- facilitates organized observation of user engagement or discussions that are open to the public;
- handles sensitive personal data, including political opinions;
- functions on an extensive level;

- impacts vulnerable data subjects, such as minors, migrants, or marginalized individuals; or
- integrates or repurposes datasets in manners that elevate inference risks.

These criteria are consistent with the guidelines on DPIAs of the form Article 29 Working Party.

Controllers, when conducting a DPIA, should:

1. Identify inherent risks and assess the likelihood and severity of harm;
2. Implement technical and organisational controls to mitigate residual risk;
3. Re-assess risk when purposes or datasets change; and
4. Record outcomes within the DPIA, including Data Protection Officer (DPO) recommendations and mitigation measures.

In cases where residual risk is still significant following mitigation actions, the Controller is required to consult the relevant Data Protection Authority prior to deployment (Art. 36 GDPR).

### Implementation steps for CEPs

- Identify the roles of each stakeholder (controllers, processors, and joint controllers) and execute the respective Data Protection Agreement, as required.
- Ensure compliance by maintaining the appropriate documentation, e.g. a Record of Processing Activities (RoPA), Data Protection Impact Assessments (DPIAs) for high-risk AI features before their launch. These documents must be updated regularly.
- Apply technical controls by enforcing role-based access, pseudonymisation, utilizing encryption, and ensuring retention practices align with defined purposes.
- Configure cookie/consent mechanisms: Activate only the essential functionalities of the platform required for participation; refrain from engaging in unnecessary tracking or profiling.
- Adhere to the principle of data minimisation, while complying with logging obligations.
- Lawful processing of user-generated content for benchmarking or training AI tools necessitates a solid public-interest justification, effective safeguards, or clear consent (Kuner, Bygrave, & Docksey, 2020).

## 6.3 Digital Services Act (DSA)

### What it is & why it matters.

The Digital Services Act (DSA; European Commission, 2024)) (Regulation (EU) 2022/2065) created a unified regulatory framework for intermediary services in the European Union's digital single market. The Regulation sets forth stricter requirements for online platforms that share and host user-generated content. Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) face the most stringent set of obligations.

CEPs that gather and disseminate feedback from citizens must adhere to the accountability and transparency standards set forth by the Digital Services Act (DSA). Clear terms of service, efficient notice-and-action processes, explanations for moderation decisions, accessible internal complaint mechanisms, and regular transparency reports are essential components.

### **Risk-based approach -duties scale-**

- The DSA utilizes a tiered framework that adjusts regulatory requirements based on the risk profile of a platform.
- All intermediary services must designate contact points for authorities and users and are required to collaborate with national Digital Services Coordinators while ensuring their terms of service are accessible (European Commission, 2024). This requirement also specifies terms that clearly outline a service's content moderation policy and the rationale behind their algorithmic systems.
- Hosting services and online platforms are required to fulfill certain obligations. These include implementing an effective notice-and-action mechanism for the removal of illegal content, providing justification for any access restrictions or content removals, establishing internal complaint-handling and out-of-court dispute resolution procedures, and publishing regular public transparency reports.
- VLOPs and VLOSEs are subject to additional controls that are proportionate to systemic risk. This encompasses risk assessments, risk-mitigation measures, independent auditing, access to data for vetted researchers, and an enhanced compliance governance model.
- This approach guarantees that regulation aligns with a platform's possible societal effects and its ability to influence public discourse.

### **Implementation steps for CEPs**

- Effective notice-and-action mechanisms. Establish straightforward and user-friendly methods for reporting illegal or inappropriate content; ensure timely acknowledgment and processing of notifications; respond with due diligence; and maintain thorough documentation of each decision and subsequent actions.
- Provide written justifications. When content is removed, demoted, or user access is restricted, it is essential to provide a statement of reasons that outlines the legal or contractual basis, the factual rationale, and the available options for redress.
- Freedom of expression and integrity – Moderation policies must strike a balance between safeguarding deliberative integrity and ensuring protections for lawful expression. To maintain user trust and legitimacy, AI-assisted moderation should be integrated with human review and clear reasoning.

## **6.4 NIS2 & CRA**

### **What they are & why they matter**

The NIS2 Directive (Directive (EU) 2022/2555) establishes a baseline for cybersecurity risk management and incident reporting across 18 sectors for both "essential" and "important" entities within the EU. It mandates that Member States oversee and enforce these responsibilities. Public administrations and civic-tech providers that integrate CEPs may be included based on their mandate, size, and criticality. Even if they are not included, the practices outlined in NIS2 - such as

governance, policies, testing, and incident processes—are considered best practices for CEPs' operations.

The Cyber Resilience Act (CRA) (Regulation (EU) 2024/2847) establishes secure-by-design mandates for "products with digital elements" (hardware/software). This includes requirements for lifecycle vulnerability management, security updates, technical documentation, and the reporting of actively exploited vulnerabilities.

### Implementation steps for CEPs

- **Treat the CEP as part of the security perimeter:** adopt risk-management measures (policies, asset lists, access control, logging/monitoring, backup/recovery, supplier security), test them regularly, and evidence effectiveness.
- **Stand up incident handling to NIS2 patterns:** implement tiered **incident notification**
- **Procurement clause check:** require vendors of CEP components to warrant NIS2 alignment (secure update policy, vulnerability handling, incident notice, technical docs), and to support audits/evidence production.

## 6.5 Accessibility frameworks: Web Accessibility Directive & European Accessibility Act

### What it is & why it matters.

Accessibility serves as a legal requirement rather than a mere optional aspect of digital civic engagement. The Web Accessibility Directive (Directive (EU) 2016/2102) mandates that public-sector websites and mobile applications must be perceivable, operable, understandable, and robust, in accordance with the harmonised EN 301 549 standard. Public entities managing CEPs are required to include an accessibility statement and a feedback system that allows users to report any barriers encountered (European Parliament & Council, 2016; ETSI, 2021).

The European Accessibility Act (Directive (EU) 2019/882) is applicable to certain consumer-facing products and services outlined in the directive. These include e-commerce services, consumer banking, electronic communications services, e-books, transport ticketing and information tools, as well as specific ICT hardware such as ATMs, payment terminals, smartphones, and e-readers. Private CEPs are included only if they offer one of the specified covered services, particularly e-commerce. Micro-enterprises that offer services are not subject to the EAA service obligations, and all operators can utilize fundamental-alteration or disproportionate-burden assessments when appropriate. Transitional rules are applicable to legacy products and services, as well as to pre-existing contracts. Public-sector CEPs are required to comply with WAD/EN 301 549, and vendors should demonstrate conformance to EN 301 549 during procurement. Additionally, EAA duties are applicable only when the CEP encompasses a covered consumer service or product (European Parliament and Council, 2019)

### Implementation steps for CEPs

- Design to EN 301 549 from the outset: integrate accessibility into specifications, budget for continuous accessibility QA, and avoid “retrofitting.”
- Support multilingualism & plain language: ensure content is readable and accessible to minority-language speakers and low-literacy groups.
- Test with assistive technologies: verify compatibility with screen readers, captioning, keyboard navigation, and voice interfaces.
- Incorporate accessibility into procurement processes by mandating that vendors provide documentation demonstrating compliance with key requirements of accessibility directives.
- Inclusion and data protection are interconnected, as the implementation of personalized assistive features frequently requires the processing of sensitive personal data. Establishing a valid legal basis for this processing is crucial, along with the implementation of suitable technical and organizational security measures.

## 7. Insights from ITHACA

This section synthesizes perspectives from citizens, practitioners (e.g. experts working with vulnerable or marginalized people), civic organizations and officials on the opportunities and risks of AI-enabled civic participation. The various stakeholders were involved in a series of participatory activities during the implementation of ITHACA project.

### 7.1 Method in brief (tools, data, limitations)

#### Approach.

ITHACA partners organised **participatory and thematic workshops**; **focus groups** (including a Delphi Study) and **expert consultations** with stakeholders in the two pilot countries of the project (Romania and Slovakia) and across Europe. Discussions explored key concerns—**trust, transparency, inclusion, accessibility, accountability, and fairness**—and expectations regarding AI features such as clustering, translation, summarisation, and content moderation. This process was complemented by comparative analysis of EU-level guidance on consultation (European Commission, 2021), disinformation (European Commission, 2022), and accessibility (ETSI, 2021).

Citizens' Representatives from the pilot countries were engaged in a thematic workshop and a focus group (Delphi Study), to identify the main facilitators (or motivating and enabling factors) and barriers (or obstacles) for the engagement of marginalized groups in online civic activities (Kocsis, 2024). Citizens' Representatives are experts working on a daily basis with socially vulnerable citizens, play a key role in representing their interests on social and/or political level, and have good knowledge in regard to the inclusion-related aspects and civic participation of socially vulnerable groups well-grounded in the local and national contexts.

Citizen' Juries were involved in two rounds of participatory workshops to explore their views, understanding, fears and motivation to engage in online civic activities. Expectations and concerns in regard to AI use in CEPs were also discussed. Citizen' Juries are two groups of people who were selected in the two cities participating as pilots in the ITHACA project, trying to achieve a high degree of diversity in regard to backgrounds and representation of socially vulnerable or marginalized citizens.

Civic organizations and experts working on design and development of CEPs and relevant AI tools (including technical experts, data analysis and ethical advocates) were engaged in focus groups (including structured interviews and roundtables) to discuss both ethical and technical considerations of AI-enabled CEPs that facilitate inclusion and accessibility.

#### Evidence base.

Inputs to the Evidence (E) base included transcripts and notes from focus groups and interviews, questionnaire answers and observations from participatory and thematic workshops, and **academic research** on digital democracy and civic technologies (e.g., Fishkin, 2018; Helbing, 2019). This triangulation allowed us to capture both lived experience and normative insights from theory and practice.

#### Limitations.

Findings are **qualitative and thematic**; they do not claim statistical representativeness. They also reflect contexts with **different digital maturities** and **policy salience**, meaning transferability

requires careful consideration. Moreover, while accessibility standards are clear, inclusion practices are highly **context-dependent**.

## 7.2 Insights from workshops, focus groups and consultations

### Transparency and traceability are decisive (E1).

Participants and experts highlighted that democratic legitimacy depends on visible **status tracking** of contributions (*received* → *under review* → *accepted/rejected* → *implemented*) and the provision of rationales (*accountability and transparency of the decision-making process*). This echoes the principle of the **right to good administration** in EU law (Charter of Fundamental Rights, art. 41) and findings from OECD deliberative-process evaluations (OECD, 2020). When traceability is absent, citizens perceive engagement as symbolic rather than consequential.

### Plain-language explanations build trust (E2).

Focus group participants strongly preferred **concise rationales** over technical or legalistic responses. Short, accessible explanations—anchored in specific criteria such as budgetary or legal limits—were perceived as respectful and legitimate. This finding is consistent with research on **public reason and deliberative democracy**, which emphasises clarity and accessibility of justification (Habermas, 1996; Fishkin, 2018).

### AI assistance must be transparent and contextual (E3).

Stakeholders supported AI use for clustering, translation, and summarisation but insisted that outputs must be labelled “**AI-assisted**” and paired with explanatory **feature cards**. Concealed automation risks undermining trust (European Parliament & Council, 2024). This aligns with broader scholarship on algorithmic accountability, which stresses the importance of **explainability** for legitimacy (Wieringa, 2020).

### Human oversight as a safeguard (E4).

Both citizens and officials emphasised that final moderation or enforcement actions must involve human judgment. AI should assist by flagging, but humans must decide, record reasons, and offer appeals. This mirrors requirements in the **AI Act** (European Parliament & Council, 2024) and principles of **due process** in administrative law (Craig, 2012).

### Process nudges vs. opinion steering (E5).

Neutral reminders about process deadlines were welcomed, but personalised prompts suggesting preferred opinions were perceived as manipulative. This reflects wider ethical concerns about **covert influence** in AI systems (Floridi & Cowls, 2019).

### AI as a mediator for digital and civic literacy (E6).

The representatives, citizens and AI experts highlighted the role of AI as mediator for personalized micro-learning (e.g. transform between formats, gamify the process, provide real-time feedback, self-paced) to facilitate technology acceptance and lead to enhanced digital and civic literacy. This aligns with the research showing that tools like chatbots, natural language processing and predictive analytics can lower accessibility barriers and facilitate skill and knowledge acquisition (Sarafis, 2025).

**Training is the key (E7).**

The participants emphasised the importance of the training data sets and continual training, as AI algorithms are only as good as the data they are trained on (e.g. bias can be introduced, contexts may evolve). This reflects the challenges of dynamic socio-technical systems, where interactions are required to adapt to highly heterogeneous contexts (user diversity, changing needs, understanding dynamic norms) (Freitas dos Santos, 2023).

## 7.3 Inclusion readiness vs accessibility

**Accessibility is a legal floor; inclusion is a practice (E8).**

Technical compliance with **Web Accessibility Directive 2016/2102** and **EN 301 549** is a baseline obligation (European Parliament & Council, 2016; ETSI, 2021). However, stakeholders stressed that inclusion requires **organisational readiness**: outreach roles, budgets, community partnerships, and support structures to enable under-represented groups to participate effectively. Research on digital divides reinforces this, showing that access gaps are less about connectivity and more about **capability and support** (van Dijk, 2020).

**Language justice beyond translation (E9).**

Machine translation lowers barriers but often erases nuance, especially in minority or regional languages. Participants emphasised the need for **plain-language drafting, community review,** and human translation of critical texts. This echoes UNESCO's call for **linguistic diversity** as a pillar of inclusive digital policy (UNESCO, 2021).

**Assisted drafting as empowerment (E10).**

Templates and structured prompts helped new participants articulate input without altering content. This is consistent with research on **participatory design**, which highlights that content-neutral scaffolding increases participation without shaping viewpoints (Carreira et al., 2022).

**The importance of last-mile support (E11).**

Under-represented groups often rely on **offline-to-online bridges**—community liaisons, assisted submission desks, or helplines. Without these supports, participation tends to skew toward already empowered actors. This aligns with studies of civic tech deployments that show success depends on **hybrid infrastructures** bridging digital and physical spaces (Peixoto & Sifry, 2017).

**Trust signals must be live, not symbolic (E12).**

Logos and compliance statements were seen as insufficient. Citizens wanted **named contact points, timelines, and “you said — we did” reports**. This resonates with the European Commission's Better Regulation principles, which stress **closing the feedback loop** as essential for legitimacy (European Commission, 2021).

## 7.4 What didn't work (and why)

- **Over-aggregation of input:** Automated clustering erased minority perspectives; aligning with OECD (2021) recommendations, minority views must be explicitly surfaced.
- **Generic moderation policies:** Rules designed for social media proved unsuitable for civic debates; proportionality and contextuality are needed.

- **Opaque AI outputs:** Summaries without provenance or source links undermined credibility (Wieringa, 2020).
- **Broad data reuse clauses:** Citizens resisted blanket consents for AI training, consistent with GDPR's **purpose limitation** principle (European Parliament & Council, 2016).
- **Notification fatigue:** Frequent alerts without clear value reduced engagement, consistent with behavioural research on digital participation (Bail, 2021).

## 7.5 Implications for policy recommendations

The evidence suggests that effective AI-enabled civic participation requires attention to **three layers**:

1. **Process quality:** visible contribution tracking, concise rationales, and timely responses are critical.
2. **Responsible AI use:** transparent labelling, explainability, human oversight, and appeals should be defaults.
3. **Inclusion readiness:** accessibility compliance must be paired with outreach, multilingualism, and offline-to-online support.

These insights provide the normative and empirical basis for the **policy recommendations in Unit 8**, ensuring they are grounded in both stakeholder voices and European legal-ethical frameworks.

## 8. Policy Recommendations for AI-enabled CEPs

### 8.1 Transparency, Accountability & Integrity

#### A. Context

Transparency and accountability are core values of any healthy democracy - and CEPs are no exception. CEPs have the potential to play an active role in public debates, rather than just neutral facilitators. They amplify some over others, influencing which opinions and issues get more visibility to decision-makers than others, summarize or rank community's discussions in particular ways. It is important to understand how these platforms operate, and to make them understandable for citizens. In situations where citizens are unable to observe the outcomes of their contributions or understand the rationale behind the removal of certain comments, the authenticity of participation may be compromised, leading to a potential loss of trust (Helberger, Pierson, & Poell, 2021).

Chapter 6 highlighted the existing legal framework of the EU, which imposes several obligations pertaining to transparency, accountability, and security. The DSA mandates that platforms disclose their moderation policies and offer justifications for the removal or limitation of content. Additionally, it requires them to notify users about the utilization of automated tools (European Parliament & Council, 2022). The AI Act establishes requirements for documentation, logging, and human oversight for systems that have a considerable social impact (European Parliament & Council, 2024). Concurrently, the NIS 2 Directive recommends that public platforms enhance their cybersecurity and data integrity measures (OECD, 2023).

Nevertheless, a gap persists between the rules of law and how they are applied in practice. Many civic-tech platforms fail to conform to legal standards and do not reflect the foundational principles of the law. As a result, a variety of challenges require attention and resolution.

#### B. Problems

- Opaque moderation systems

Automated moderation systems frequently render decisions without transparency regarding their processes or rationale. The ambiguity surrounding whether a human or an algorithm has responded to user posts creates uncertainty, resulting in frustration and diminishing trust in the civic participation process.

- Absence of contestability

The DSA requires statements of reasons; however, these often lack specificity or completeness. Users are unable to effectively appeal a decision or comprehend the standards being utilized without clear explanations (Gorwa, Binns, & Katzenbach, 2020).

- Insufficient auditability and oversight

Due to the proprietary nature of many moderation systems, independent experts and regulators face challenges in verifying their fairness and accuracy. (Gorwa, Binns, & Katzenbach, 2020).

- Algorithmic bias and its associated chilling effects

AI moderation often misinterprets specific languages, tones, or expressions from minority groups. This results in an imbalanced environment where certain communities are unjustly muted, undermining the diversity of participation essential for democratic engagement (Matamoros-Fernández & Farkas, 2021).

- Risks associated with integrity and manipulation

CEPs face manipulation tactics akin to those observed on social media, including bot attacks, coordinated disinformation efforts, and mass reporting campaigns. Such threats have the potential to disrupt civic dialogue and undermine the legitimacy of public consultation (Council of Europe, 2023).

## C. Recommendations

- Transparent moderation policies

CEPs should make their moderation rules transparent and comprehensible to everyone. Policies must explain what counts as acceptable or prohibited content, where AI tools are involved, and when human review is applied. Updates and examples should be published in all relevant languages to make the process predictable and inclusive (Helberger et al., 2021).

- Statements of reasons and appeals

Each Moderation Action ought to be accompanied by a concise and precise explanation that directly references the specific rule that has been applied. It is essential that citizens possess the right to contest a decision via an internal review process and, if necessary, through a secondary appeal body. This straightforward procedural safeguard contributes to the transformation of online participation into a process characterized by fairness and accountability (Gorwa et al., 2020).

- Registrar of AI systems and audit logs

Each platform should keep a public register of the AI systems it uses (including GAI systems/models) — their purpose, data sources, and last update. Moderation decisions should be logged and anonymised so they can be reviewed by oversight bodies. These measures turn the abstract principle of transparency into something concrete and verifiable (Binns, 2020).

- Independent audits and oversight

In the context of policy making where the use of CEPs is mandatory, it is essential that regular audits conducted by external experts be established as a standard practice. The purpose of these audits is to assess the consistency of content moderation practices, monitor for bias, and ensure that systems comply with users' rights. Sharing summaries of findings in clear language would help the public understand that oversight is meaningful and not just symbolic (Dryzek et al., 2019, Matamoros-Fernández & Farkas, 2021; OECD, 2023).

- Integrity and resilience safeguards

Considering the various incidents that have been identified as attempts to manipulate elections in the digital realm, safeguarding civic spaces online necessitates protection against such manipulation. CEPs ought to implement principles of cybersecurity-by-design in accordance with NIS 2. The publication of an annual integrity report detailing these initiatives would serve to illustrate accountability and enhance credibility.

- Clear allocation of responsibilities

Transparency is effective only when responsibility is both shared and clearly defined. The platform provider must ensure documentation and algorithmic transparency; while the platform operator (for instance, a municipality) must guarantee lawful data processing and user rights.

| <b>Recommendation</b>                         | <b>Responsible Actor(s)</b>                | <b>Core Action</b>   | <b>Relevant Legal / Policy Framework</b>                      |
|---|--|--|---|
| <b>Transparent moderation policies</b>        | CEP operator (e.g. municipality, ministry) | Draft and publish moderation policy including AI-use thresholds, rules, and examples in plain, multilingual format.          | DSA (EU) 2022/2065; AI Act (EU) 2024/1689; GDPR (EU) 679/2016 |
| <b>Statements of reasons and appeals</b>      | CEP operator                               | Implement standardised Statement of Reasons (SoR) templates and establish a two-tier appeals process.                        | DSA   |
| <b>Registrar of AI systems and audit logs</b> | AI system provider                         | Maintain and disclose a registry of AI systems (purpose, dataset, version); retain anonymised moderation logs for oversight. | AI Act; DSA   |
| <b>Independent audits and oversight</b>       | Lawmakers / regulators                     | Establish obligation for independent audits; publication of plain-language summaries.  | AI Act  |
| <b>Integrity and resilience safeguards</b>    | AI system provider & CEP operator          | Apply NIS2-aligned cybersecurity-by-design measures; issue annual integrity report.  | NIS2 Directive (EU) 2022/2555; Council of Europe Rec(2023)    |

## 8.2 Privacy, Data Governance & Open Data

### A. Context

CEPs necessitate the processing of personal data, including special categories of personal data, which are vital for the operation of democratic processes. In user-generated content, such as posts, comments, or proposals, individuals may disclose information regarding a person's political beliefs, background, vulnerability status, sexual orientation, family life, or geographical location. Unlawful processing of this information may significantly erode trust in the overall participatory process (Bennett & Oduro-Marfo, 2019). Furthermore, certain characteristics of CEPs, including participation in consultative referendums, may necessitate the collection of substantial identification data during the registration process. In this context, privacy and protection of personal data go beyond mere adherence to regulatory framework; it serves as an essential element for safeguarding individuals within the participation process, enabling individuals to express their opinions freely without the fear of surveillance or unethical profiling.

For CEPs, the challenge is to strike the right balance: citizens must be confident that their personal information is protected, while society should still benefit from aggregated, anonymised insights that make governance more accountable and evidence-based. This balance between individual rights and collective transparency is at the core of responsible data governance for civic participation.

### B. Problems

- Excessive collection of personal data and profiling

Many CEPs collect more personal data than is necessary - from demographic surveys to behavioural analytics. The collection of personal data breaches the GDPR's principle of data minimisation and risks creating profiles that could be used for political or commercial targeting (Mantelero, 2018).

- Inconsistent risk assessment

While DPIAs are mandatory for high-risk data processing activities under GDPR, they are often conducted retrospectively or superficially. Public authorities are still unaware of FRIAs, introduced under the AI Act, leaving cumulative risks to fundamental rights unaddressed (Mantelero, A. 2024).

- Chilling effects on participation

When individuals believe that their contributions could be monitored, repurposed, or examined for purposes not directly related to their input, they may choose to refrain from sharing their thoughts or disengage entirely from the process. The apprehension surrounding surveillance notably deters vulnerable or minority groups from participating (Bennett & Oduro-Marfo, 2019).

- Lack of open and standardised data for accountability

Although anonymised open data could help evaluate fairness and policy impact, many CEPs keep those data locked in proprietary systems. When datasets are shared, they often lack proper documentation or anonymisation (Janssen & Helbig, 2021).

- Insufficient governance of institutional data

Municipalities and other public authorities frequently lack the expertise and resources to handle data effectively. Without a clear data governance plan, practices regarding data management vary widely, leading to fragmented protection across Europe (OECD, 2023).

## C. Recommendations

- Data minimisation and lawful basis mapping

The organizations acting as data controllers when operating CEPs, such as municipalities and NGOs, need to make sure they collect only the personal data that is essential for fulfilling their purpose -namely, to improve civic participation through the platform's features. This is in line with the idea of data minimization as described under Article 5(1)(c) GDPR. All types of personal data collected through the CEP needs to be connected to a specific legal basis as outlined in Article 6 of the GDPR. It is crucial to ensure that the processing is essential for the intended purpose. When processing depends on consent, it is essential to make sure that consent is freely given, informed, and can be easily withdrawn. It is important to emphasize that individuals do not need to consent in order to exercise their right to participate in formal decision-making processes in public institutions, as this right is based on Article 6(1)(e) (public task/exercise of official authority). CEPs must by default use anonymization and/or pseudonymization procedures, especially when the data is meant to be reused in analytics or research.

- Integrated DPIA–FRIA process

Prior to the launch of the platform or a new AI feature, the CEP operator —such as a municipality, a ministry, or an NGO - should carry out a DPIA alongside a FRIA in a cohesive manner. DPIAs primarily focus on privacy issues, while FRIAs adopt a broader perspective by examining the social implications of AI, including issues of discrimination, fairness, and exclusion. Although DPIAs and FRIAs are not always required, conducting them is highly recommended to enhance the protection of citizens' fundamental rights through their involvement in civic participation processes. Policymakers must consider the implications of mandating FRIA for all CEPs in future planning. This measure would ensure a consistent approach to protecting rights.

- Clear and accessible privacy notices

Privacy notices ought to be clear, structured, and expressed in straightforward, easy-to-understand language. The CEP operator shall explain to the users what types of data is collected, why it is needed, how long it will be stored, and who will have access. Contextual “just-in-time” messages integrated within the platform can assist users in grasping privacy implications right when they engage with it. Making information accessible and available in multiple languages is crucial for fostering inclusiveness.

- DPO Appointment

Each CEP operator must appoint a Data Protection Officer (DPO) as required by Article 37 GDPR. The DPO should oversee DPIAs and FRIAs, verify anonymisation methods, and act as the contact point for data subjects and supervisory authorities. National or regional agencies could support smaller authorities by providing standardised templates, training, and anonymisation guidelines, fostering more uniform compliance across jurisdictions (OECD, 2023).

- Open Data

Sharing open data using anonymized datasets can improve public accountability and aid academic research, as long as strong safeguards are implemented. According to the Regulation on European Data Governance (EU) 2022/868 (Data Governance Act/DGA) and the Directive (EU) 2019/1024 on open data, the re-use of public sector information should adhere to established anonymisation standards, ensuring that both direct and indirect identifiers are eliminated. Tiered access systems, which differentiate between completely open datasets and those that require controlled access for

accredited researchers, can effectively balance the need for transparency with the necessity of protection (Janssen & Helbig, 2021).

| Recommendation  | Responsible Actor(s)                                | Core Action  | Relevant Legal / Policy Framework                   |
|---|---|--|---|
| <b>Data minimization and lawful basis mapping</b>                 | CEP operator (Data Controller)                      | Identify lawful bases for all processing activities; maintain RoPA; apply pseudonymisation and retention limits. | GDPR  |
| <b>Integrated DPIA–FRIA process</b>                               | CEP operator (Data Controller/Deployer) & Lawmakers | Conduct combined DPIA–FRIA before AI deployment; lawmakers to consider mandating FRIA for all CEPs that use AI.  | GDPR; AI Act; EDPB Guidelines                       |
| <b>Clear and accessible privacy notices</b>                       | CEP operator (Data Controller)/ DPO                 | Draft layered, multilingual privacy notices; embed just-in-time explanations in interfaces.                      | GDPR; Web Accessibility Directive (EU) 2016/2102    |
| <b>Institutional data-governance capacity and DPO appointment</b> | CEP operator (Data Controller)                      | Appoint DPO; provide templates, training, and periodic internal privacy audits.                                  | GDPR; OECD Digital Government Recommendation (2023) |
| <b>Anonymised open data publication</b>                           | CEP operator (Data Controller)                      | Publish anonymised datasets under DGA/ODD with documentation and tiered access controls.                         | Data Governance Act; Directive on open data         |

## 8.3 Responsible & Explainable AI for Deliberation

### A. Context

AI can help improve CEPs. AI can assist the operators of CEPs to process vast quantities of information, whether that is clustering proposals, synthesizing citizens’ comments, or detecting toxic content. AI features can be applied to participatory processes to make them more inclusive and comprehensible (Zangl et al., 2025); therefore, CEPs have the potential to strengthen civic engagement and encourage citizens’ involvement in democratic participatory processes.

Conversely, AI presents several new challenges. The AI Act highlights that some systems are considered high-risk in light of their potential impact on society and sets out numerous requirements in relation to risk management, transparency, and human oversight (European Parliament & Council, 2024). The Council of Europe (2023) likewise states that any algorithmic system deployed for democratic life shall enhance – not replace – the human capacity for judgment and free expression. For CEPs, this entails that the automated tools support constructive deliberation rather than being an impediment to it.

## B. Problems

- Opacity and lack of explainability

Many AI-assisted features used for summarisation or recommendation work as “black boxes.” Citizens and even administrators often cannot tell how the system arrived at a particular clustering or conclusion, which undermines both interpretability and trust (Zangl et al., 2025).

- Over-reliance on automation

Automated text summarisation or sentiment analysis can unintentionally simplify complex arguments. When relied upon without human review, these tools risk flattening dissent and turning deliberation into a numbers game rather than a thoughtful exchange (Carnegie Endowment, 2025).

- Weak risk governance

Few public operators maintain systematic AI risk registers or conduct recurring bias audits. Without continuous monitoring, small technical errors can become systemic distortions that influence public decisions unnoticed (OECD, 2023).

- Limited contestability for users

Citizens rarely have an avenue to question how their contributions were summarised or ranked. The absence of feedback mechanisms leaves them feeling unheard, which weakens procedural fairness and perceived legitimacy (Binns, 2020).

## C. Recommendations

- AI risk management and accountability

Every CEP should maintain a detailed AI risk classification register listing each system’s purpose, training-data origin, AI-assisted outputs, known limitations, and potential risks to the health, safety, or fundamental rights of individuals. Depending on risk categorization, specific compliance measures shall be implemented. Assigned professionals should periodically review these registers, ensuring that risk documentation becomes a living compliance instrument rather than a one-time exercise (European Parliament & Council, 2024).

- Human-in-the-loop oversight

AI must never be the final arbiter in matters that affect participation or human rights. Human reviewers should validate any automated outcome that could exclude, classify, or prioritise citizens' input. Clear escalation procedures should exist so that ambiguous or disputed cases are resolved by moderators trained in both ethics and deliberative standards (Helberger et al., 2021).

- Explainable AI and user transparency

AI-driven decisions should be interpretable for both administrators and participants. CEPs can highlight the text segments or reasoning patterns that influenced an output, provide confidence scores, and offer “Why am I seeing this?” explanations. Such visibility makes algorithmic reasoning accessible to citizens and reinforces democratic accountability (Cheong, Filimon, & Cole, 2024).

- Preserving deliberative quality and pluralism

AI-generated summaries must remain traceable to the original contributions and include a balanced reflection of majority, minority, and dissenting views. CEP operators should measure deliberative quality not only by participation volume but by fairness, inclusiveness, and perceived legitimacy (Dryzek et al., 2019). These qualitative indicators turn technology from a filtering mechanism into a facilitator of genuine dialogue.

| Recommendation  | Responsible Actor(s)                   | Core Action  | Relevant Legal / Policy Framework   |
|---|--|--|---|
| <b>AI risk management and accountability structures</b>     | Platform operator                      | Maintain AI risk register and implement the compliance project according to AI categorization of CEPs.             | AI Act  |
| <b>Human-in-the-loop oversight</b>                          | Platform operator                      | Establish human review for any automated decision with rights impact; define escalation and override protocols.    | AI Act, DSA, GDPR   |
| <b>Explainable and transparent interactions and AI</b>      | AI System Provider                     | Integrate Explainable AI (XAI) features: reasoning highlights, confidence scores, “Why am I seeing this?” options. | AI Act, GDPR  |
| <b>Preservation of deliberative integrity and pluralism</b> | AI System Provider & Platform operator | Ensure AI-generated summaries reflect diverse viewpoints; link summaries to original inputs.                       | AI Act, Council of Europe Rec(2023); OECD (2020) Guidelines on Deliberation |

## 8.4 Inclusion, Accessibility & Deliberative Quality

### A. Context

Inclusion is what makes democratic participation meaningful. CEPs have the potential to support people to make their voice heard, but they can only do so if designed and operated in ways that remove physical, language and cognitive barriers. Accessibility is therefore not only a legal requirement, but also a moral commitment to the principle of equal citizenship (W3C/WAI, 2018).

The Web Accessibility Directive (Directive (EU) 2016/2102) and the European Accessibility Act (Directive (EU) 2019/882) oblige public authorities and service providers to ensure that digital services are perceivable, operable, understandable and robust for all users. Technical specifications such as EN 301 549 translate these requirements into verifiable design criteria. But accessibility is not an end in itself. Inclusion is about a diversity of voices, constructive debate and responsive institutions as a prerequisite for democratic participation (OECD, 2023; Dryzek et al., 2019).

CEPs have a dual role to play. It is important not only to make sure that all people have access, but also to enable deliberation – that is, that all voices, whether majority or not, can be expressed, heard and fairly considered in the decision-making process.

### B. Problems

- Ongoing gaps in digital access.

Access to technology varies significantly depending on factors like age, income, and location. In the absence of parallel offline channels or focused support, CEPs may unintentionally widen existing inequalities instead of bridging them (van Dijk, 2020).

- Challenges related to language and literacy.

Many platforms tend to miss the mark when it comes to offering multilingual interfaces or straightforward language options. This oversight can leave out individuals who either don't speak the main language or struggle with digital literacy (Spiliopoulou et al., 2020).

- Barriers to accessibility.

Even with clear guidelines in place, numerous civic platforms struggle to fully meet accessibility standards. Individuals with disabilities often encounter forms that are not accessible or lack proper compatibility with screen readers (W3C/WAI, 2018).

### C. Recommendations

- Accessibility compliance and hybrid participation

All CEPs operated by or for public authorities should comply fully with the Web Accessibility Directive, the European Accessibility Act, and EN 301 549. Accessibility must cover web and mobile interfaces, ensuring perceivability and usability for people with diverse abilities. To prevent exclusion of those without reliable internet access, authorities should maintain equivalent offline participation channels—such as postal submissions, telephone surveys, or community workshops—mirroring online functions (European Parliament & Council, 2019).

- Multilingual and plain-language design

Interfaces, notifications, and explanatory texts should be available in all official and regionally relevant languages. Real-time translation, combined with human moderation, helps avoid semantic distortion in deliberative contexts. Content should follow plain-language principles and readability testing, supporting citizens with low literacy or cognitive impairments (Spiliopoulou et al., 2020).

- Targeted outreach and inclusion programs

Authorities and platform operators should actively reach out to groups under-represented in civic life—older adults, migrants, low-income citizens, or rural residents. Measures could include community facilitators drawn from local networks, incentives such as transport vouchers or childcare support, and collaboration with NGOs trusted by the target groups. Inclusion efforts should be evaluated periodically using demographic and geographic participation indicators (Friess & Eilders, 2015).

| Recommendation   | Responsible Actor(s) | Core Action  | Relevant Legal / Policy Framework  |
|--|----------------------|--|--|
| <b>Accessibility compliance and hybrid participation</b> | Platform provider    | Ensure compliance across web and mobile applications; provide equivalent offline participation channels. | Web Accessibility Directive (EU) 2016/2102; European Accessibility Act (EU) 2019/882 |
| <b>Multilingual and plain-language design</b>            | Platform provider    | Offer multilingual interfaces and plain-language content; integrate translation tools.                   | Charter Art. 21  |

|   |                                      |  |   |  |
|---|--------------------------------------|--|---|--|
| <p><b>Targeted and inclusion measures</b></p> | <p><b>inclusion and outreach</b></p> | <p>CEP operator (Municipality / NGO)</p> | <p>Develop outreach programs for under-represented groups; partner with local NGOs; evaluate inclusiveness.</p> | <p>OECD Recommendation on Open Government (2017); Council of Europe Rec(2023)1</p> |
|---|--------------------------------------|--|---|--|

## 8.5 Institutional Integration, Traceability & Platform Governance

### A. Context

For CEPs to have democratic value, participatory processes should be institutionally embedded within public decision-making rather than conducted as temporary pilot initiatives or communication activities. Institutional anchoring of participation processes ensures that they are not sporadic, but rather integrated into regular governance cycles and that citizens’ inputs are not easily ignored or marginalised (Fung, 2015). This allows them to connect with policy outcomes and enable affected citizens to track how their contributions have been considered or how they contributed to final decisions.

In this respect, it is worth noting that several public authorities, including the European Commission with its Better Regulation Agenda, already have mechanisms for public consultation, feedback, or issuance of synopsis reports detailing how citizens’ input was taken into account (European Commission, 2021). However, these requirements are not consistently and homogeneously applied by all EU member states and local authorities. The OECD (2023) or the Council of Europe (2023), among others, also highlight that democratic innovation should move from experimentation to institutionalisation, so that participatory processes are no longer exceptional, but part of standard democratic governance.

This also relates to platform governance: the way CEPs are procured, managed, operated and maintained, as well as who has control over them. Public authorities that rely on services provided by vendors with proprietary systems have limited auditability and control in the long term. Institutional integration of participatory processes, therefore, also concerns the technology and tools that support them. It is a question of maintaining the autonomy of the public sphere to ensure that civic infrastructure does not depend on privately held interests and values.

### B. Problems

- Ad hoc use of CEPs

Many CEPs are launched as part of time-bound projects or consultations with no legal or policy requirements to keep them operational after the funding expires. Platforms and related institutional memory are then often abandoned after a project ends, a status quo event or consultation (Smith, 2009).

- Lack of traceability and policy linkage

Citizen contributions are sometimes lost in admin workflows with no clear linkage to policy outputs, limiting the public's ability to trace how and when their input was used. This lack of transparency discourages participation and erodes democratic accountability (OECD, 2023).

- Vendor lock-in and lack of transparency

Public bodies may become locked into proprietary civic-tech vendors and frameworks without the ability to fully access data, algorithms, and interoperability specifications due to market dominance or exclusion of open source options in procurement. This creates high exit costs and makes it harder to meet transparency requirements under EU law (van Dijck, Poell, & de Waal2018).

- Weak procurement safeguards

Procurement contracts for CEPs and AI-assisted features often do not include clauses related to auditability, explainability, transparency, or rights to ensure oversight and avoid vendor lock-in. This leaves authorities unable to monitor vendor practices vis-à-vis democratic principles or legal requirements (Ada Lovelace Institute et al., 2021).

- Lack of institutional capacity

Smaller municipalities or NGOs may lack staff with training and experience on how to design effective digital participation processes or oversee data governance and AI integration. This uneven capacity creates gaps in democratic safeguards across Europe (OECD, 2023).

- Limited representativeness

While people who are already highly educated or engaged in politics are more likely to participate, certain underrepresented communities may continue to be largely absent from CEPs. This leads to selection bias and “echo chambers” rather than a deliberative “melting pot” of diverse voices (Friess & Eilders, 2015).

- Lack of consistent support from institutions

Institutions that facilitate and oversee the process can guarantee equitable terms of engagement and empower participants; but, in certain instances, the outcomes of an otherwise inclusive process fail to influence decision-making procedures. Consultations lacking feedback mechanisms or mandatory participation may be perceived as superficial (Michels, 2011).

## C. Recommendations

- Embedding participation in governance cycles

Make CEP-based consultation phases mandatory across legislative, regulatory, budgeting, and spatial planning procedures, with narrowly framed emergency derogations. Codification should specify triggers (e.g. threshold of policy impact), minimum consultation windows, and required outputs (synopsis reports). (TEU Art. 10(3); TFEU Art. 11; European Commission Better Regulation Agenda2021).

- Deliberative quality evaluation & duty to respond

Introduce a legal duty to respond: every consultation ends with a plain-language “You said – We did” report linking inputs (or clusters) to final measures (accepted/modified/rejected) and reasons. Evaluate deliberative quality using mixed methods (fairness, representativeness, inclusiveness, and policy-impact indicators) and publish the indicator set ex ante. Tie compliance to decision approval (i.e., no final adoption without the response report). (OECD, 2020; Council of Europe Rec(2023)1; Better Regulation Toolbox).

- Traceability-by-design

Require CEPs to generate machine-readable audit trails: contribution IDs, versioning, timestamps, moderator actions and exportable logs. Define retention, access control, and role-based permissions; The traceability schema should integrate with records management and transparency obligations (GDPR 2016/679; DSA; AI Act 2024/1689).

- Transparent procurement

Embed enforceable clauses in all CEP/AI contracts: algorithmic transparency and explainability (incl. documentation access), third-party audit rights, data portability, interoperability via open standards, security-by-design, logging, and exit terms to avoid vendor lock-in. Use model clauses and make them public (AI Act 2024/1689; Data Act 2023/2854; NIS2 2022/2555).

- Public-interest foundations & open-source civic tech

Establish/finance public-interest entities to maintain open-source CEP frameworks and vetted AI-assisted features (moderation, translation, summarisation) with published conformance profiles and security hardening guides. Provide long-term maintenance, threat-monitoring, and compliance updates; allow smaller authorities to onboard with minimal local capacity. Encourage shared component audits and pooled bug-bounty schemes. (OECD, 2023; Open Data Directive 2019/1024).

- Institutional learning & capacity building

Create continuous training pathways for officials on participatory design, algorithmic accountability, DPIA/FRIA practice, accessibility, and rights-based procurement. (OECD, 2023).

| Recommendation   | Responsible Actor(s)                  | Core Action  | Relevant Legal / Policy Framework  |
|--|---------------------------------------|--|--|
| <b>Embedding participation in governance cycles</b>          | Lawmakers                             | Codify mandatory CEP-based consultation phases in legislative, regulatory, budgeting and planning procedures; allow only narrow emergency derogations  | TEU Art. 10(3); TFEU Art. 11; European Commission Better Regulation Agenda (2021)                              |
| <b>Deliberative quality evaluation &amp; duty to respond</b> | Public authorities                    | Evaluate fairness, representativeness, and policy impact using mixed methods; publish plain-language “You said – We did” reports linking inputs to outcomes (accepted/modified/rejected, with reasons) | OECD (2020) Guidelines on Deliberative Processes; Council of Europe Rec(2023)1; Better Regulation Toolbox      |
| <b>Traceability-by-design</b>                                | Platform provider & Platform operator | Implement contribution IDs, versioning, machine-readable audit trails, and linkage to the relevant administrative dossier; define retention and export for oversight/research                          | GDPR (EU) 2016/679—accountability; DSA (EU) 2022/2065—transparency; AI Act (EU) 2024/1689—logging & oversight  |
| <b>Transparent &amp; rights-based procurement</b>            | CEP operators (Contracting Agencies)  | Include enforceable clauses for explainability, audit rights, data portability, interoperability (open standards), security-by-design, logging, and exit to avoid vendor lock-in                       | Directive 2014/24/EU (Public Procurement); AI Act (EU) 2024/1689; Data Act (EU) 2023/2854; NIS2 (EU) 2022/2555 |

| Recommendation  | Responsible Actor(s)                               | Core Action  | Relevant Legal / Policy Framework                          |
|---|--|--|--|
| <b>Public-interest foundations &amp; open-source civic tech</b> | Member States & European Commission                | Establish/finance foundations maintaining OSS CEP frameworks and vetted AI modules (moderation, translation, summarisation); publish conformance profiles and compliance guides  | OECD (2023) Digital Government; Open Data Directive;       |
| <b>Institutional learning &amp; capacity building</b>           | Public Authorities; National Schools of Government | Continuous training on participatory governance, algorithmic accountability, DPIA/FRIA practice, and rights-based procurement; EU-supported peer-learning hubs and model clauses | OECD (2023) Digital Government; EDPB DPIA guidance; AI Act |

## 9. Conclusion

As democratic life becomes increasingly mediated through digital infrastructures, civic participation must evolve to reflect this reality. CEPs enriched by AI capabilities, present a powerful opportunity to broaden access to public deliberation, streamline complex consultations, and reinforce institutional responsiveness. Yet, these same technologies carry risks that, if left unaddressed, can undermine transparency, fairness, and trust in democratic processes. This White Paper sets out a principled and practical response to this duality, advancing a coherent framework for policy, governance, and design of AI-enabled civic participation across the EU.

This White Paper provides a structured set of policy recommendations anchored in five thematic pillars: (1) Transparency and Democratic Accountability; (2) Privacy, Data Governance, and Open Data; (3) Responsible and Explainable AI for Deliberation; (4) Inclusion, Accessibility, and Deliberative Quality; and (5) Institutional Integration and Platform Governance. More than a conceptual guide, this White Paper is designed as an actionable policy instrument. It translates ethical and legal principles into concrete governance measures to ensure that AI supports—not replaces—public reasoning. It also provides role-specific guidance to legislators, municipalities, civic tech providers, and oversight bodies, clarifying shared responsibilities in building lawful, inclusive, and trustworthy digital participation infrastructures.

Finally, the conclusions and recommendations of this White Paper will directly inform the development of the “CEPs Data Governance Framework”, which will operationalize these values into technical, procedural, and legal standards. Together, these deliverables articulate a unified vision for embedding civic participation into the digital transformation of democratic institutions—one grounded in rights, transparency, and the imperative of civic empowerment.

## 10. References

1. **European Parliament & Council of the European Union. (2012).** Charter of Fundamental Rights of the European Union (2012/C 326/02). <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX%3AC2012%2F326%2F02> EUR-Lex
2. **European Parliament & Council of the European Union. (2002).** Directive 2002/58/EC (ePrivacy Directive). <https://eur-lex.europa.eu/eli/dir/2002/58/oj/eng> EUR-Lex
3. **European Parliament & Council of the European Union. (2016).** Regulation (EU) 2016/679 (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/req/2016/679/oj/eng> EUR-Lex
4. **European Parliament & Council of the European Union. (2016).** Directive (EU) 2016/2102 on the accessibility of the websites and mobile applications of public sector bodies. <https://eur-lex.europa.eu/eli/dir/2016/2102/oj/eng> EUR-Lex
5. **European Parliament & Council of the European Union. (2019).** Directive (EU) 2019/882 on the accessibility requirements for products and services (European Accessibility Act). Official Journal of the European Union, L 151, 70–115. <https://eur-lex.europa.eu/eli/dir/2019/882/oj/eng>
6. **European Parliament & Council. (2019).** Directive (EU) 2019/1024 ... (Open Data Directive). Official Journal of the European Union, L 172.
7. **European Parliament & Council. (2022b).** Regulation (EU) 2022/868 ... (Data Governance Act). Official Journal of the European Union, L 152.
8. **European Parliament & Council of the European Union. (2022).** Regulation (EU) 2022/2065 (Digital Services Act). <https://eur-lex.europa.eu/eli/req/2022/2065/oj/eng> EUR-Lex
9. **European Parliament & Council of the European Union. (2022).** Directive (EU) 2022/2555 (NIS 2 Directive). <https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng> EUR-Lex
10. **European Parliament & Council of the European Union. (2024).** Regulation (EU) 2024/1689 (Artificial Intelligence Act). <https://eur-lex.europa.eu/eli/req/2024/1689/oj/eng> EUR-Lex
11. **European Commission. (2020).** European democracy action plan (COM(2020) 790). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52020DC0790> EUR-Lex
12. **European Commission. (2021).** Better Regulation Guidelines (SWD(2021) 305). [https://commission.europa.eu/system/files/2021-11/swd2021\\_305\\_en.pdf](https://commission.europa.eu/system/files/2021-11/swd2021_305_en.pdf) European Commission
13. **European Commission. (2021).** Better regulation: Joining forces to make better laws. Brussels: European Commission. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021DC0219>
14. **European Commission. (2021).** Proposal for a Regulation laying down harmonised rules on artificial intelligence. <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence>
15. **European Commission. (2023, July 20).** Better regulation: Guidelines and toolbox. [https://commission.europa.eu/law/law-making-process/better-regulation/better-regulation-guidelines-and-toolbox\\_en](https://commission.europa.eu/law/law-making-process/better-regulation/better-regulation-guidelines-and-toolbox_en) European Commission
16. **European Commission. (2025, July 10).** The General-Purpose AI Code of Practice. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
17. **European Commission. (2025).** Cyber Resilience Act. <https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act> Ψηφιακή Στρατηγική Ευρώπης
18. **European Commission. (2022).** The 2022 Code of Practice on Disinformation. <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation> Ψηφιακή Στρατηγική Ευρώπης
19. **European Commission / AccessibleEU. (2025, Jan 31).** The EAA comes into effect in June

2025. Are you ready? [https://accessible-eu-centre.ec.europa.eu/content-corner/news/eea-comes-effect-june-2025-are-you-ready-2025-01-31\\_en](https://accessible-eu-centre.ec.europa.eu/content-corner/news/eea-comes-effect-june-2025-are-you-ready-2025-01-31_en) AccessibleEU

**20. European Data Protection Supervisor (EDPS). (2024).** Generative AI and the EUDPR. First EDPS Orientations for EUIs using Generative AI. [https://www.edps.europa.eu/system/files/2024-06/24-06-03\\_genai\\_orientations\\_en\\_0.pdf](https://www.edps.europa.eu/system/files/2024-06/24-06-03_genai_orientations_en_0.pdf)

**21. European High Level Expert Group on AI (AI HLEG). (2019).** Ethics guidelines for trustworthy AI. [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419)

**22. ETSI. (2021).** EN 301 549 v3.2.1: Accessibility requirements for ICT products and services. [https://www.etsi.org/deliver/etsi\\_en/301500\\_301599/301549/03.02.01\\_60/en\\_301549v030201p.pdf](https://www.etsi.org/deliver/etsi_en/301500_301599/301549/03.02.01_60/en_301549v030201p.pdf)  
ETSI

**23. ENISA. (2025, June).** Technical implementation guidance on cybersecurity risk-management measures (NIS2). [https://www.enisa.europa.eu/sites/default/files/2025-06/ENISA\\_Technical\\_implementation\\_guidance\\_on\\_cybersecurity\\_risk\\_management\\_measures\\_version\\_1.0.pdf](https://www.enisa.europa.eu/sites/default/files/2025-06/ENISA_Technical_implementation_guidance_on_cybersecurity_risk_management_measures_version_1.0.pdf)

**24. W3C/WAI. (2018).** Web Content Accessibility Guidelines (WCAG) 2.1. World Wide Web Consortium. <https://www.w3.org/TR/WCAG21/>

**25. OECD. (2020).** Good practice principles for deliberative processes for public decision making. <https://www.oecd.org/content/dam/oecd/en/topics/policy-issue-focus/innovative-citizen-participation/good-practice-principles-for-deliberative-processes-for-public-decision-making.pdf>  
OECD

**26. OECD. (2021).** Evaluation guidelines for representative deliberative processes. [https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/11/evaluation-guidelines-for-representative-deliberative-processes\\_10b0cea1/10ccbfcfb-en.pdf](https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/11/evaluation-guidelines-for-representative-deliberative-processes_10b0cea1/10ccbfcfb-en.pdf) OECD

**27. OECD. (2022).** OECD Framework for the Classification of AI Systems: A tool for effective AI policies. <https://read.oecd.org/10.1787/cb6d9eca-en?format=pdf>

**28. OECD. (2023).** Government at a Glance 2023. <https://doi.org/10.1787/3d5c5d31-en> OECD

**29. OECD. (2020).** Catching the deliberative wave: Innovative citizen participation and new democratic institutions. Paris: OECD Publishing. <https://doi.org/10.1787/339306da-en>

**30. UNESCO. (2021).** Recommendation on the ethics of artificial intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000385082> unesdoc.unesco.org

**31. Council of Europe. (2018).** Recommendation CM/Rec(2018)4 on the participation of citizens in local public life. <https://rm.coe.int/16807954c3> rm.coe.int

**32. Council of Europe. (2023).** Report on deliberative democracy and the risks of digital manipulation. Strasbourg: Council of Europe. <https://rm.coe.int/cddg-2023-22e-report-of-the-18th-meeting-2771-8516-7625-v-1/1680addfb7>

**33. Council of Europe. (2024).** Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). <https://rm.coe.int/1680afae3c> rm.coe.int

**34. ADL. (2019).** Online hate and harassment: The American experience | ADL. <https://www.adl.org/resources/report/online-hate-and-harassment-american-experience>

**35. Ada Lovelace Institute, AI Now Institute, & Open Government Partnership. (2021).** Algorithmic accountability for the public sector: Learning from impact assessments worldwide. London: Ada Lovelace Institute.

**36. Carnegie Endowment for International Peace. (2025).** How AI can unlock public wisdom and revitalize democratic governance. Washington, DC: Carnegie Endowment. <https://carnegieendowment.org/posts/2025/07/how-ai-can-unlock-public-wisdom-and-revitalize->

[democratic-governance?lang=en](#)

- 37. World Economic Forum. (2025, October).** Cybersecurity and information integrity: Managing hybrid threats to democracy. <https://www.weforum.org/stories/2025/10/cybersecurity-information-integrity>
- 38. Bennett, C. J., & Oduro-Marfo, S. (2019).** Privacy, voter surveillance and democratic engagement: Challenges for data protection authorities. Paper presented at the 41st International Conference of Data Protection and Privacy Commissioners (ICDPPC), Tirana. [https://globalprivacyassembly.com/wp-content/uploads/2019/10/Privacy-and-International-Democratic-Engagement\\_finalv2.pdf](https://globalprivacyassembly.com/wp-content/uploads/2019/10/Privacy-and-International-Democratic-Engagement_finalv2.pdf)
- 39. Abdelhalim, E., Anazodo, K. S., Gali, N., & Robson, K. (2024).** A framework of diversity, equity, and inclusion safeguards for chatbots. *Business Horizons*, 67(5), 487–498.
- 40. Achara, S., & Chhabra, M. (2025).** Bias, fragility and safety challenges in AI content moderation systems. arXiv preprint arXiv:2501.13302. <https://arxiv.org/abs/2501.13302>
- 41. Akbarighatar, P., Pappas, I. O., & Vassilakopoulou, P. (2023).** Justice as fairness: A Hierarchical framework of responsible AI principles. *ECIS 2023 Research-in-Progress Papers*, 79. [https://aisel.aisnet.org/ecis2023\\_rip/79](https://aisel.aisnet.org/ecis2023_rip/79)
- 42. Bail, C. (2021).** *Breaking the social media prism: How to make our platforms less polarizing.* Princeton University Press.
- 43. Binns, R. (2020, January).** On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAcCT/FAT\*)* (pp. 514–524). <https://doi.org/10.1145/3351095.3372864>
- 44. Carreira, P., Ferreira, J., & Silva, A. (2022).** Designing inclusive civic tech: A participatory design perspective. *Journal of Community Informatics*, 18(1), 55–72.
- 45. Cheong, M., Filimon, A., & Cole, S. (2024).** Transparency and accountability in AI: A multidimensional framework. *Frontiers in Human Dynamics*, 6, 1421273. <https://doi.org/10.3389/fhumd.2024.1421273>
- 46. Craig, P. (2012).** *Administrative law* (7th ed.). Oxford University Press.
- 47. De Gregorio, G. (2020).** Democratising online content moderation: A constitutional framework. *Computer Law & Security Review*, 36, 105374.
- 48. Dryzek, J. S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J. N., Felicetti, A., ... & Warren, M. E. (2019).** The crisis of democracy and the science of deliberation. *Science*, 363(6432), 1144–1146. <https://doi.org/10.1126/science.aaw2694>
- 49. Dylan, H., & Grossfeld, E. (2025).** Revisionist future: Russia’s assault on large language models, the distortion of collective memory, and the politics of eternity. *Dialogues on Digital Society*. <https://doi.org/10.1177/29768640251377941>
- 50. ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018).** Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).
- 51. ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021).** Latent hatred: A benchmark for understanding implicit hate speech. In M. F. Moens et al. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 345–363). <https://doi.org/10.48550/arXiv.2109.05322>
- 52. Falco, E., & Kleinhans, R. (2018).** Beyond technology: Identifying local government challenges for using digital platforms for citizen engagement. *International Journal of Information Management*, 40, 17–20. <https://doi.org/10.1016/j.ijinfomgt.2018.01.007>
- 53. Farina, C. R. (2014).** Designing an online civic engagement platform. Cornell Law Faculty Publications, Paper 124. <https://scholarship.law.cornell.edu/facpub/124>

- 54. Fishkin, J. S. (2018).** Democracy when the people are thinking: Revitalizing our politics through public deliberation. Oxford University Press.  
<https://www.amphilsoc.org/sites/default/files/2020-03/attachments/Fishkin.pdf>
- 55. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020).** Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, 2020(1).
- 56. Floridi, L., & Cows, 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>
- 57. Freitas dos Santos, T., Osman, N., & Schorlemmer, M. (2023).** A multi-scenario approach to continuously learn and understand norm violations. Auton Agent Multi-Agent Syst 37, 38.  
<https://doi.org/10.1007/s10458-023-09619-4>
- 58. Friess, D., & Eilders, C. (2015).** A systematic review of online deliberation research. Policy & Internet, 7(3), 319–339. <https://doi.org/10.1002/poi3.95>
- 59. Fung, A. (2015).** Putting the public back into governance: The challenges of citizen participation and its future. Public Administration Review, 75(4), 513–522.  
<https://doi.org/10.1111/puar.12361>
- 61. Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012).** Social media use for news and individuals' social capital, civic engagement and political participation. Journal of Computer-Mediated Communication, 17(3), 319–336. <https://doi.org/10.1111/j.1083-6101.2012.01574.x>
- 62. Gorwa, R., Binns, R., & Katzenbach, C. (2020).** Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 1–15.  
<https://doi.org/10.1177/2053951719897945>
- 63. Goyal, P., Huang, Y., Wang, S., Taly, A., & Chi, E. H. (2025).** MoMoE: Mixture of Moderation Experts for large-scale content moderation. arXiv preprint arXiv:2505.14483.  
<https://arxiv.org/abs/2505.14483>
- 64. Habermas, J. (1996).** Between facts and norms: Contributions to a discourse theory of law and democracy. MIT Press.
- 65. Hagendorff, T. (2020).** The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120.
- 66. Hagendorff, T. (2022).** Blind spots in AI. AI and Ethics, 2, 851–867.
- 67. Hagendorff, T. (2024).** Mapping the ethics of generative AI: A comprehensive scoping review. Minds and Machines, 34(4), 39.
- 68. Halevy, M., Harris, C., Bruckman, A., Yang, D., & Howard, A. (2021).** Mitigating racial biases in toxic language detection with an equity-based ensemble framework. In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1–11).
- 69. Helberger, N., Pierson, J., & Poell, T. (2017).** Governing online platforms: From contested to cooperative responsibility. The Information Society, 34(1), 1–14.  
<https://doi.org/10.1080/01972243.2017.1391913>
- 70. Helbing, D. (2019).** Towards digital enlightenment: Essays on the dark and light sides of the digital revolution. Springer.
- 71. Janssen, M., & Helbig, N. (2021).** Innovating and leveraging data for public governance: Building data infrastructures for accountability and participation. Information Polity, 26(4), 475–488.  
<https://doi.org/10.1016/j.qiq.2015.11.009>
- 72. Jhaver, S., Bruckman, A., & Gilbert, E. (2019).** Does transparency in moderation really matter? User behavior after content removal explanations on reddit. Proceedings of the ACM on Human-Computer Interaction, 3, 1–27.
- 73. Jobin, A., Ienca, M., & Vayena, E. (2019).** The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

- 74. Kocsis, O., de Lera, E., Bedek, M. A., Zangl, M., Colt, R., Sausova, V., & Birasova, L. (2024).** Facilitators and Barriers of Marginalized Groups for the Engagement in Online Civic Activities. *International Conference on Electronic Government and the Information Systems Perspective*, 105–121. [https://doi.org/10.1007/978-3-031-68211-7\\_9](https://doi.org/10.1007/978-3-031-68211-7_9)
- 75. Kuner, C., Bygrave, L. A., & Docksey, C. (Eds.). (2020).** *The EU General Data Protection Regulation (GDPR): A commentary*. Oxford University Press.
- 76. Luckett, J. (2023).** Regulating generative AI: a pathway to ethical and responsible implementation. *Journal of Computing Sciences in Colleges*, 39(3), 47–65.
- 77. Mantelero, A. (2018).** AI and big data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4), 754–772. <https://doi.org/10.1016/j.clsr.2018.05.017>
- 78. Mantelero, A. (2024).** The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template. *Computer Law & Security Review*, 54, 105045. <https://doi.org/10.1016/j.clsr.2024.106020>
- 79. Martínez-Gil, J., Sanz, I., Miguélez, J., & Díez, A. (2025).** An overview of civic engagement tools for rural communities. *Open Research Europe*, 4, 195. <https://doi.org/10.12688/openreseurope.16818.1>
- 80. Matamoros-Fernández, A., & Farkas, J. (2021).** Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2), 205–224. <https://doi.org/10.1177/1527476420982230>
- 81. Michels, A. (2011).** Innovations in democratic governance: How does citizen participation contribute to a better democracy? *International Review of Administrative Sciences*, 77(2), 275–293. <https://doi.org/10.1177/0020852311399851>
- 82. Molina, M. D., & Sundar, S. S. (2022).** When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4), 1–12.
- 83. Pavlopoulos, J., Sorensen, J., Laugier, L., & Androutsopoulos, I. (2021).** SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)* (pp. 59–69).
- 84. Peixoto, T., & Sifry, M. L. (2017).** *Civic tech in the global south: Assessing technology for the public good*. World Bank Publications.
- 85. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August).** "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).
- 86. Sarafis, D., Karamitsios, K., & Kravari, K. (2025).** AI and Civic Engagement: A Brief Exploration of Applications and Opportunities. *Proceedings of International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, 1–6. DOI: 10.1109/ICADEIS65852.2025.10933183
- 87. Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021).** AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection. *IEEE Transactions on Technology and Society*, 2(1), 31–42. <https://doi.org/10.1109/TTS.2021.3052127>
- 88. Smith, G. (2009).** *Democratic innovations: Designing institutions for citizen participation*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511609848>
- 89. Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019).** What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13, 1526–1543.
- 90. Voigt, P., & Von dem Bussche, A. (2017).** *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer.

- 91. van Dijk, J. (2020).** The digital divide. Cambridge: Polity Press. DOI:10.1002/asi.24355
- 92. van Dijck, J., Poell, T., & de Waal, M. (2018).** The platform society: Public values in a connective world. Oxford University Press. DOI:10.23860/MGDR-2018-03-03-08
- 93. Waseem, Z., Davidson, T., Warmesley, D., & Weber, I. (2017).** Understanding abuse: A typology of abusive language detection subtasks. In 1st Workshop on Abusive Language Online (ALW 2017) (pp. 78–84). Association for Computational Linguistics (ACL).
- 94. Weisz, J. D., He, J., Muller, M., Hofer, G., Miles, R., & Geyer, W. (2024).** Design Principles for Generative AI Applications. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1–22).
- 95. Wieringa, M. (2020).** What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18.
- 96. Zangl, M., Loi, I., Zachos, P., Bedek, M., Dimogerontakis, E., Nikolaou, C.-E., & Moustakas, K. (2025).** A multidisciplinary analysis of transparent AI-driven toxicity-detection tools for civic engagement platforms. AI & Society. <https://doi.org/10.1007/s00146-025-02424-5>