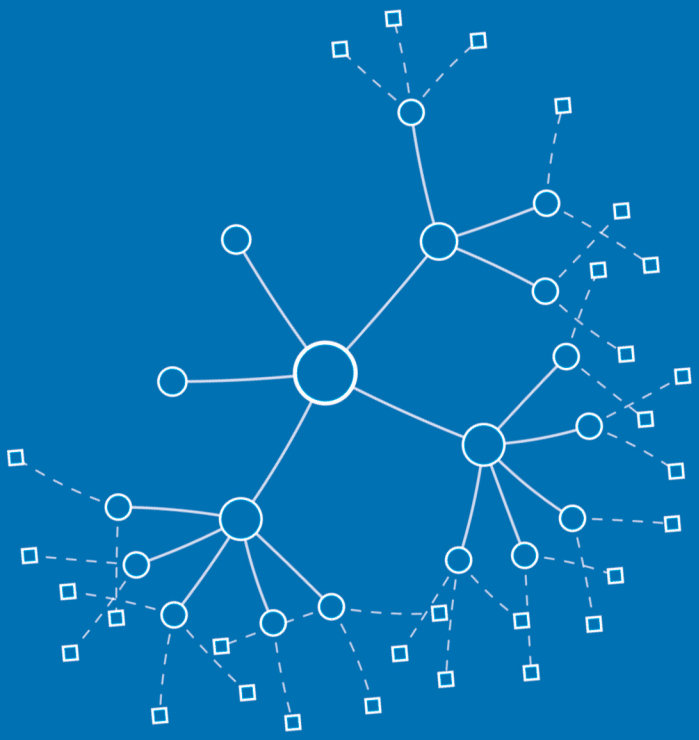
From Knowledge Graphs to Civic Sensemaking

Communication Design for Human-Centered Explainable AI in AI-Enhanced Deliberation



Politecnico di Milano | School of Desing
MsC in Communication Design
Supervisor: Ilaira Mariani
Giacomo Garetto [247607]
A.Y. 2024-25

Communication Design
School of Design
Politecnico di Milano
AY 2024-25

Discussant: Giacomo Garetto
Supervisor: Ilaria Mariani

INDEX

TABLE OF FIGURES

TABLE OF TABLES

ABSTRACT

AI is increasingly embedded in civic technologies to moderate, structure and summarize large-scale public discussions. While these capabilities can scale participation, they also introduce an interpretability gap: the transformations performed by AI mediation pipelines (e.g., clustering, summarization, ranking) often remain hard to understand, verify and contest for heterogeneous publics, with potential consequences for democratic legitimacy.

This thesis reframes explainability in civic deliberation as a communication design problem. It combines a systematic literature review across XAI, human-centered XAI, information visualization, data storytelling and civic tech, with a Research Through Design process conducted within the EU ORBIS project. Insights from the review are translated into a consolidated set of requirements and design principles, aligned with ORBIS constraints, and operationalized in Il Corollario—an interactive knowledge-graph interface that represents positions, arguments, AI-generated clusters and keywords, and supports progressive disclosure, traceability to source contributions and uncertainty framing.

The prototype is integrated into the ORBIS reporting ecosystem and evaluated through user feedback and a final expert validation against the requirements. Results indicate that combining network visualization with narrative and interaction scaffolds can reduce cognitive load, improve orientation and support critical engagement with AI outputs, while highlighting remaining risks of misinterpretation and the need for clearer cues and verification pathways.

The work contributes (1) a requirements-driven framework for communication design in AI-supported deliberation, and (2) a validated design artifact showing how interface-level mediation can support intelligibility and contestability of AI-assisted collective representations.

L'AI viene sempre più integrata nelle tecnologie applicate al contesto civico per moderare, strutturare e sintetizzare discussioni pubbliche su larga scala. Sebbene queste capacità possano aumentare la partecipazione, introducono anche un divario di interpretabilità: le trasformazioni eseguite dai processi di mediazione dell'intelligenza artificiale (ad esempio clustering, sintesi e classificazione) rimangono spesso difficili da comprendere, verificare e contestare per un pubblico eterogeneo, con potenziali implicazioni per la legittimità democratica.

Questa tesi riformula la spiegabilità nella deliberazione civica come problema di communication design. Il lavoro combina una systematic literature review su XAI, Human-Centered XAI, information visualization, data storytelling e civic tech con un percorso di Research Through Design svolto nel contesto del progetto europeo ORBIS. Le evidenze della review vengono tradotte in un set consolidato di requisiti e principi di progetto, coerenti con vincoli e obiettivi di ORBIS, e implementate in Il Corollario: un'interfaccia basata su knowledge graph che rappresenta posizioni, argomenti, cluster e keyword generate dall'AI e abilita la rivelazione progressiva dei contenuti, tracciabilità alle fonti ed inquadramento dell'incertezza.

Il prototipo è integrato nell'ecosistema di reportistica ORBIS e valutato tramite feedback degli utenti e una validazione finale con esperti, allineata ai requisiti. I risultati suggeriscono che l'integrazione di network visualization con strutture narrative e di interazione può ridurre il carico cognitivo, migliorare l'orientamento e sostenere un ingaggio critico verso gli output dell'AI, evidenziando al contempo rischi residui di errata interpretazione e la necessità di indicazioni e percorsi di verifica più chiari ed espliciti.

Il lavoro contribuisce con (1) un framework di requisiti per il design della comunicazione applicato alla deliberazione AI-enhanced e (2) un artefatto progettuale validato che mostra come la mediazione a livello di interfaccia possa supportare intelligibilità e contestabilità delle rappresentazioni collettive assistite dall'AI.

1.0
INTRODUCTION. AI-MEDIATED DELIBERATION AND THE CHALLENGE OF UNDERSTANDABILITY

1.1 AI in Civic Tech and Democratic Deliberation

The increasing integration of Artificial Intelligence (AI) into civic technologies (Civic Tech) is reshaping contemporary practices of democratic deliberation (Carstens & Friess, 2024). Within the broader field of Civic Tech, AI systems are being adopted to support participatory process at scale, responding to challenges related to inclusivity, responsiveness, quality of online deliberation and the large volumes of citizen-generated inputs (Buhmann & Fieseler, 2023; Grancea & Tutui, 2025; Vagnoni & Palmirani, 2025; Zeginis et al., 2026). These developments mark a shift in the role of digital technologies in governance, positioning AI as an infrastructural component that actively mediates how public discourse is collected, structured and rendered accessible and actionable.

In deliberative context, AI is increasingly deployed to facilitate interaction both beyond citizens and between citizens with public institutions through applications such as synthesis, automated translation, sentiment analysis, topic modelling, and the identification of key arguments emerging from the public debate. Natural Language Processing (NLP) techniques are used to process vast amounts of textual contributions that would otherwise exceed the cognitive capabilities and organisational capacity of traditional participatory mechanisms (Zeginis et al., 2026). However, the same conditions that make digital participation scalable, also intensify well-documented problems of contemporary public communication: fragmentation of publics, uneven visibility, toxic interactions, misinformation and high moderation costs. These tensions are often framed through the lens of deliberative democracy, which provides a normative vocabulary for what democratic communication should achieve, and consequently what civic tech systems are expected to support.

Deliberative democratic theory broadly holds that collective decisions gain legitimacy when they are preceded by public reasoning among free and equal participants (Gutmann & Thompson, 2004). While deliberative scholarship contains multiple models and internal debates, a shared common core is often expressed in terms of reason-giving, reciprocity and respect under conditions that approximate equality of participation (Barvosa, 2019). In online context, these ideals translate into operational expectations about discourse quality, interactional tone and participatory conditions such as inclusion and balanced voice.

For this thesis, deliberative democracy is used here primarily as a framing device rather than a full theoretical foundation: it helps justify why civic tech systems are evaluated not only in terms of usability or engagement, but also in terms of democratic qualities such as inclusiveness, reciprocity, respect and the possibility of accountable influence on outcomes.

Online spaces were initially seen by some as infrastructure for expanded public deliberation, yet empirical trajectories of large-scale online discussion have been marked by persistent challenges. Carstens and Friess (2024) summarize this tension by contrasting early optimism about deliberative potential with the observed prevalence of incivility and “irrationality” in many online discussions, which diverge from deliberative ideals. These dynamics are not simply “noise” around otherwise functional participation. In practice, they shape who speaks, whose contribution remains visible, what counts as a valid contribution, and whether participation is experienced safe and worth sustaining, particularly for marginalized groups.

Asynchronous forums, and consultation platforms must often manage large volumes of citizen input, sustain basic norms of interaction, and produce intelligible outputs for both participants and institutional decision-makers. Zeginis et al. (2026) note that online participation platforms, while promising for mass deliberation, repeatedly fall short due to low participation, opaque processes and low quality contributions—conditions that directly undermine deliberative ambitions at scale.

Against this backdrop, AI is increasingly positioned as a set of enabling capabilities for scaling, moderating and synthesizing deliberation—especially in text-based, asynchronous settings that resemble forums and comment-based consultations. A key contribution of Zeginis et al. (2026) is to structure AI support around stages of the deliberative process (preparatory, deliberation, post-processing) and to map concrete “AI features” onto these stages. Across the literature they review, AI (including LLM-based approaches) is proposed or implemented to address three recurrent bottlenecks:

- **Moderation and interaction management during deliberation.**
  Proposed features include automated moderation and turn-taking support; detection of toxic behaviour and harmful content; and facilitation mechanisms such as identifying stalls in live conversations. These functions aim to lower the costs of moderation and reduce the dominance of few “louder”

participants, thereby supporting more equitable participation conditions.

- **Reducing information overload through synthesis and structuring.**
  Post-processing features include summarization, topic modeling, clustering and organizing arguments, and aggregation across deliberation groups, transforming large volumes of unstructured discourse into manageable representations for citizens and policymakers.

- **Accessibility and inclusion in multilingual, heterogeneous publics.**
  Real time translation and text simplification may broaden participation by reducing language barriers and comprehension demands.

Fig.01 Deliberation phases compared to AI Integration possibilities

DELIBERATION PHASES	AI INTEGRATION
Live deliberation	▪ Turn-taking support ▪ Toxic behavior detection ▪ Identify Stalls ▪ Reduce dominance of “loud speakers” ▪ Real time translation ▪ Text simplification
Post-processing	▪ Summarization ▪ Topic modelling ▪ Clustering ▪ Aggregation across Groups ▪ Argument organization

This feature-oriented view clarifies why AI is increasingly attractive to civic tech: it promises a partial response to scale problems (volumes, time, moderation capacity) and to heterogeneity problems (language diversity, varied literacy levels, uneven familiarity with policy topics). At the same time, it raises a literacy challenge that is often underestimated: as deliberative processes become mediated by clustering, summarization, ranking systems and confidence framings, participation increasingly depends on citizens’ ability to interpret data visualizations, grasp the implications of algorithmic processing and recognize framing effects. These interpretative capacities are unevenly distributed across publics, potentially raising entry barriers where formal access remains open.

AI can increase efficiency but may simultaneously reduce citizen empowerment if participation becomes less intelligible, less contestable or more dependent on machine mediation. The criterion is not whether AI can improve deliberation in the general, but whether it does so while preserving meaningful human agency and accountable influence over outcomes.

This point resonates with Buhmann and Fieseler’s (2023) deliberative governance perspective on responsible AI innovation. They argue that responsible innovation in AI calls for well-informed public deliberation across public, private and civil society sectors, but that this is undermined by AI opacity and knowledge boundaries between experts and citizens—conditions that erode trust and the very possibility of deliberation about AI systems. In civic tech, the implication is direct: when AI systems mediate participation, citizens require not only access to participation channels but also intelligible accounts of how those mediations shape the process and its outputs.

While AI is often introduced with the promise of “better deliberation”, Carstens and Friess (2024) provide a targeted ethical critique of how deliberative ideals are typically translated into technical objectives. They show that AI tools for online discussion disproportionately focus on the deliberative norms of rationality and civility, frequently via simplified operationalizations such as argument mining (for rationality) and the detection of verbal markers for incivility. They argue that, when examining through a sociotechnical frame, these operationalizations can reproduce social hierarchies: privileging communication styles associated with higher education and potentially excluding minority voices whose rhetorical forms, terminology or modes of participation may not align with what the system recognize as “rational” or “civil”.

This critique is particularly consequential for public-sector civic-tech, where inclusiveness is not aspirational but integral to democratic legitimacy. If AI primarily enforces formal discourse norms that are unevenly distributed across populations, then “improving deliberation” may come at the cost of narrowing who can participate effectively and whose contributions are rendered legible. Carstens and Friess (2024) therefore argue that evaluating AI interventions beyond input-output bias metrics and toward sociotechnical impacts, including predictionable exclusion effects and the distribution of burdens.

These strands motivate the central concern that the following sections will develop: the challenges of understandability in AI-mediated deliberation. In civic-tech, AI becomes part of the public interface of democracy. As Zeginis et al. (2026) note, AI integration raises risks related to fairness and transparency alongside its potential benefits. Buhmann and Fieseler (2023) emphasize that opacity and knowledge boundaries undermine trust and deliberative conditions, making the governance of AI a deliberative problem in its own right. And Carstens and Friess (2024) show that even seemingly “pro-social” AI interventions (e.g., civility enforcement) can enact normative choices with equal effects unless their assumptions and impacts are made visible and contestable.

This thesis argues that addressing these tensions requires shifting the focus from model transparency alone to the intelligibility of AI-mediated deliberation pipelines as they are encountered through public-facing interfaces. The analytical unit is therefore the mediation pipeline (filtering, clustering, summarization, ranking and related transformations) together with its communicative representation (narrative framing, visualization, and interaction affordances) through which diverse publics can interpret, evaluate, and contest how collective outputs are produced. This framing sets up Section 1.2 by positioning interpretability and explainability as preconditions for legitimate AI-supported public discussions.

1.2 The Interpretability Gap in AI-Supported Public Discussions

As civic-tech platforms increasingly incorporate AI to moderate, structure and summarize large-scale public discussion, a central tension emerges: the very computational techniques that make deliberation scalable can also make it harder for participants to understand how collective outcomes are being produced. In practice, AI is introduced to address information overload, low-quality contributions, difficult moderation, particularly in

mass online deliberation settings where asynchronous forums and social-discussions spaces generate heterogeneous, high-volume data (Zeginis et al., 2026).
Yet these same functions such as summarization, clustering, “high potential ideas” identification, automated moderation, and related forms of content processing, interpose an additional interpretative layer between citizens’ contributions and what becomes visible, legible, and actionable within the deliberative process.

This section frames the tension as an interpretability gap: a mismatch between what AI-mediated deliberation systems computationally do to public discourse (filtering, ranking, aggregating, classifying, synthesising) and what diverse publics and institutional actors need in order to comprehend, evaluate and contest these operations as legitimate components of democratic discussion. Rather than being simply technical, this gap is socio-technical and epistemic: it concerns how AI transforms the condition under which participants can follow reasoning, recognize disagreement, judge credibility, and understand how their input relates to outputs that may shape facilitation or decision making.

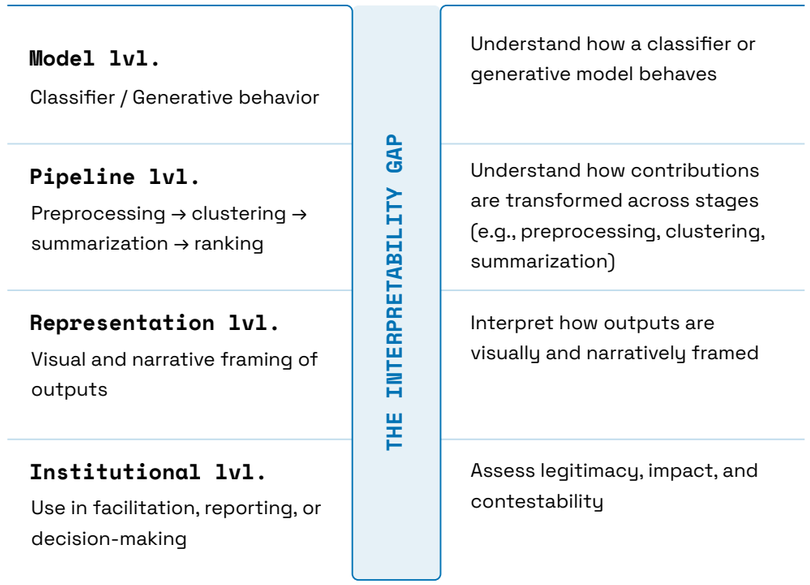


Fig.02 The interpretability gap operates across multiple levels of mediation. Rather than being simply technical, this gap is socio-technical and epistemic.

The interpretability gap operates across multiple levels of mediation. At a model level, questions concern how a classifier or generative model behaves; at a pipeline level, how contributions are transformed across stages (e.g., preprocessing, clustering, summarization); at a representation level, how outputs are visually and narratively framed for interpretation; and at an institutional level, how these representations are used in facilitation, reporting, or decision-making. While model-centered explainability typically focuses on rendering model behavior intelligible, the interpretability gap identified here concerns this broader sociotechnical mediation through which AI reshapes public discourse into public-facing representations.

In civic deliberation, this gap is consequential because AI outputs become part of the public record through which participants orient, evaluate, and contest collective representations. For this reason, the thesis treats explainability as a public-facing requirement: what matters is whether people can interpret how discourse has been transformed and what the resulting artifacts can legitimately support. Chapter 2 defines this shift conceptually, and Chapter 6 tests it through a design artifact.

1.3 Communication Design as a Mediating Practice between AI Systems and Publics

As AI becomes embedded in civic technologies for online deliberation, the democratic stakes of representation become as consequential as those of computation. In this thesis, communication design is understood as the structuring of interpretive conditions under which AI-mediated representations enter public reasoning. It operates through narrative sequencing, visual encoding, and interaction scripting, shaping how collective outputs are understood, questioned, and acted upon. In such settings, communication design can be framed as a mediating practice: it translates algorithmically produced representations into intelligible, navigable and contestable forms that heterogeneous publics can use for public reasoning. This mediation role is not aesthetic; it concerns the epistemic conditions under which participants can interpret claims, evaluate evidence and find insights and patterns in a context that is increasingly mediated by data-intensive infrastructures.

A critical starting point is recognizing that civic dashboards, AI summarise and analytic interface do not simply represent reality but they filter and structure it. Visualization scholarship argues that visualizations mediate

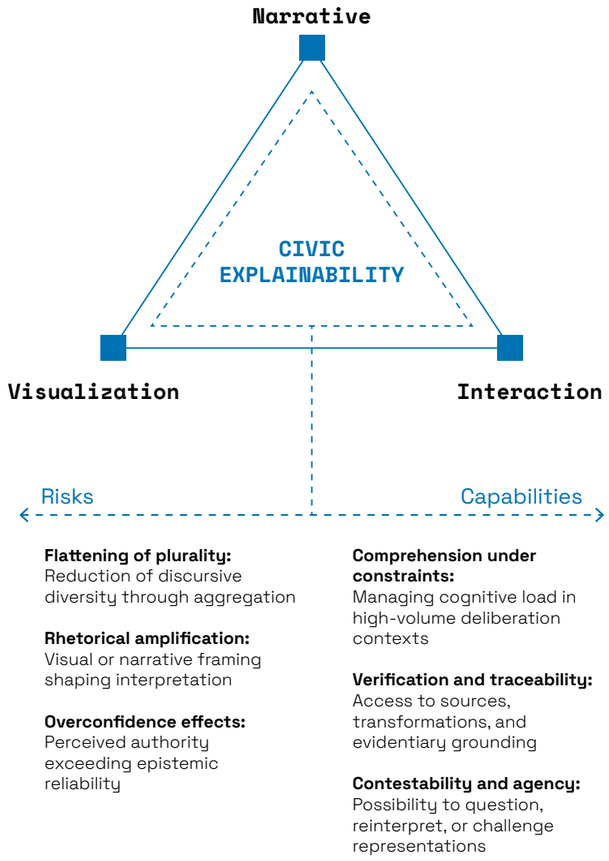
meaning through multiple transofrmations—from real world data to image and from image to interpretation—therby embedding assumptions about categories, relevance and legitimacy (Dörk et al., 2013). In AI-supported deliberation, these transformations are intensified: computational reduction is often necessary for scale, but it can also flatten pluralism by privileging dominant themes, smoothing disagreement or obscuring minority perspectives (Boehnert, 2015).

This implies that the communication design approach towards civic deliberation should treat explainability not as a model-level property, but as a public communication problem. While technical XAI methods focus on rendering model behavior intelligible, the challenge identified in Section 1.2 concerns the broader sociotechnical mediation through which AI transforms discourse into public representations. Citizens need to understand not only outputs, but also how their contributions are represented, grouped and summarized and with what limitations—so they can meaningfully evaluate, contest and reinterpret the system's mediation.

Public-sector adoption of AI is increasingly accompanied by normative and regulatory expectations concerning transparency and explainability. Herrera (2025) emphasizes that explainability is repeatedly positioned as a core principle in trustworthy/ethical AI guidelines and governance frameworks, and the degree and form of explainability required are context-dependent, especially where consequences of errors are significant. For deliberative settings, this shifts the question from "how do we expose the model?" to "how do we enable people to interpret and critically engage with AI-mediated representations in ways that support democratic reasoning?". Explainability becomes operational as intelligibility: the extent to which participants can grasp what is being represented, why it appears as it does and what uncertainties or assumptions structure it. Because interpretive capacities are unevenly distributed, this mediation must scaffold heterogeneous literacy profiles, stage complexity progressively, and enable multiple entry points into AI-mediated representations.

Operationally, this mediating layer is enacted through three coupled instruments—narrative, visualization, and interaction—that scaffold understanding without collapsing plurality and that enable inspection and contestation of AI-mediated artifacts. These instruments are developed as a framework in Chapter 5 and implemented in Il Corollario (Chapter 6).

Fig.03 Communication design as the interface between mediation pipeline and democratic interpretation.



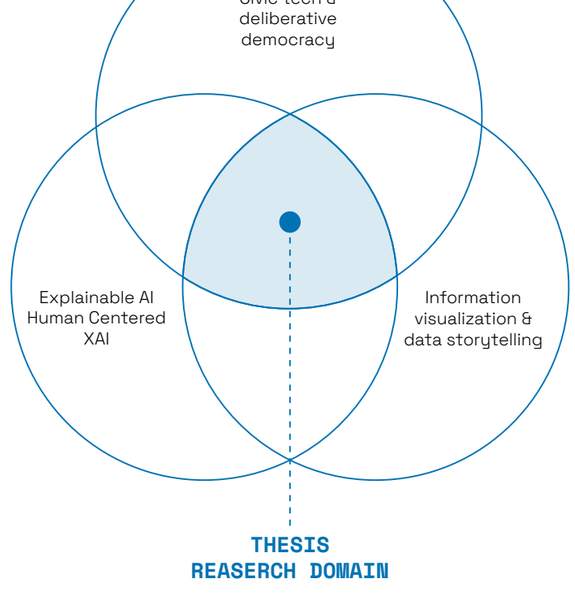
1.4 Research Domains and Approach

This thesis sits at the intersection of three research domains: (1) civic tech and deliberative democracy, as a normative frame for evaluating digital participation systems; (2) explainable AI, with a specific focus on Human-Centered XAI (HCXAI) and the limits of model-centred “transparency” in public-facing contexts; (3) information visualization and data storytelling, as mechanisms to reduce interpretive burden and support heterogeneous publics.
Rather than treating explainability as an internal property of AI models, the thesis frames explainability as a communicative condition of participation: citizens and institutions need to understand how AI-mediated transformations (e.g., clustering, summarization, ranking, annotation) shape

what becomes visible and actionable in a deliberative process, and they need interface-level means to verify and contest those transformations. The core design problem is therefore the intelligibility of the mediation pipeline as encountered through public-facing representations.

Methodologically, the work follows a Research Through Design approach that combines evidence synthesis with real-life experimentation. A systematic literature review is conducted to map the state of research at the intersection of XAI, visualization, and civic participation, and to extract design-relevant insights through structured coding and thematic synthesis. These insights are then translated into design requirements and principles, aligned with the constraints and objectives of the EU ORBIS project. The design contribution is instantiated in an interactive knowledge-graph interface (Il Corollario) integrated within the ORBIS reporting ecosystem. Finally, the thesis evaluates the resulting system through two complementary moments: user-test feedback and a final expert validation against the consolidated set of requirements, using task-based assessment and qualitative feedback to identify strengths, limitations, and design refinements.

Fig.04 Theoretical positioning of the research, the 3 fields intersection.



1.5 Structure of the Thesis

Chapter 2 defines the conceptual ground by introducing XAI and its limits in complex, public-facing settings, then positioning Human-Centered XAI as a shift from model transparency to human interpretability needs. It concludes by framing explainability as a communication/design problem and consolidating the research questions.

Chapter 3 presents the Research Through Design methodology, combining a systematic literature review (search strategy, selection, coding, synthesis) with an applied study setting. It also introduces the ORBIS project context, the BCause deliberation platform, the materials used in the study, and the iterative design process through which requirements were progressively translated into interface mechanisms.

Chapter 4 reports the results of the systematic literature review, outlining the literature landscape, disciplinary contributions, the main themes/meta-themes emerging from the synthesis, and the resulting gaps and design opportunities.

Chapter 5 translates the review evidence into a communication design framework for HCXAI in civic deliberation. It formalizes design requirements and principles and articulates “communication design as a mediating layer” through the coupled instruments of narrative, visualization, and interaction, establishing the bridge from principles to platform implementation.

Chapter 6 presents the Il Corollario interactive knowledge graph as the core design artifact: the data model and graph structure, the visual encoding and interaction rationale, and the way the interface supports collective sensemaking while keeping AI-mediated layers legible, verifiable, and contestable. It also describes how the prototype is integrated into the ORBIS reporting ecosystem.

Chapter 7 details the validation strategy and results, distinguishing between user interviews (supporting continuous improvement) and expert interviews (final requirement-based validation). It synthesizes key findings and translates them into prioritized implications for design refinement.

Chapter 8 discusses the research contributions and their implications, including what the work adds to HCXAI in civic contexts, how communication design operates as mediation, and the methodological contributions and limits of the study.

Chapter 9 concludes by summarizing implications for civic tech and AI governance and outlining future research directions.

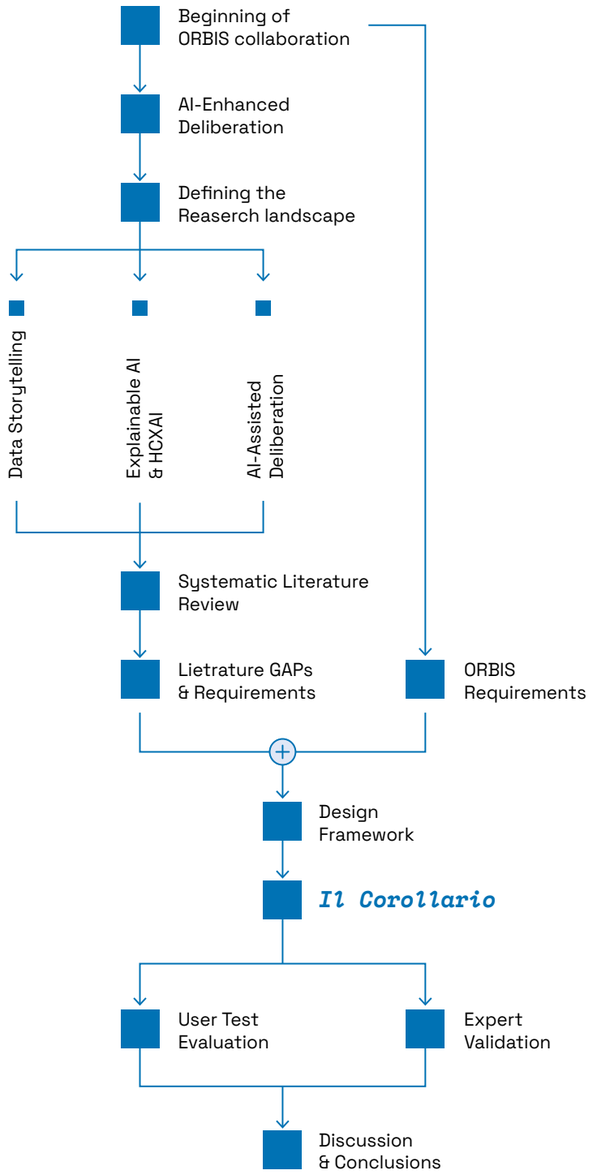


Fig.05 Thesis roadmap structure

2.0
SETTING THE GROUND.
HUMAN-CENTERED XAI

2.1 From Transparent Systems to Opaque Models. AI Explainability as model-centred logic and its limits

Before discussing explainable AI (XAI), it is useful to clarify what is meant by algorithm and why contemporary AI systems can be experienced as "opaque." Historically, an algorithm is not a mysterious entity but a method: a finite sequence of well-specified instructions that, if executed in the correct order, produces a solution to a given problem (Knuth, 1997). In everyday discourse, however, "the algorithm" is often invoked as a shorthand for an unseen decision process. This ambiguity becomes more consequential when the focus of control shifts from explicitly programmed rules to learned statistical models, where the procedure that produces an output is no longer authored line-by-line by a programmer but is instead inferred from data.

A concise historical trajectory helps situate why explainability emerges as a critical issue. Early AI research (roughly from the 1950s to the 1980s) was dominated by symbolic approaches, later framed as GOFAI (Good Old-Fashioned Artificial Intelligence) (Haugeland, 1985). In these systems, intelligence was modelled as a top-down manipulation of symbols via explicit, human-authored rules. In such systems, reasoning traces were, at least in principle, inspectable: one could follow the chain of "if-then" rules that produced an outcome. At the same time, they were brittle: minor ambiguities or unanticipated cases could cause failure, since what was not explicitly represented could not be handled robustly.

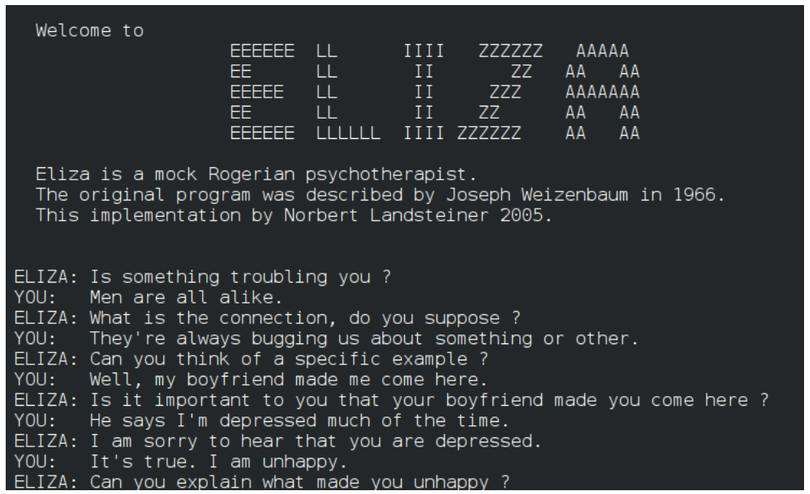


Fig.06 Example of a Conversation with the ELIZA Chatbot

retrieved from: https://commons.wikimedia.org/wiki/File:ELIZA_conversation.png

From a human–computer interaction standpoint, an important observation often underemphasized, is that opacity is not only a property of models, but also of human perception and interaction. Weizenbaum's (1966, 1976) ELIZA, the first instance of a chatbot using a digital representation of a psychologist for the purpose of parody, illustrate this point clearly: despite being a comparatively simple pattern-matching script, it elicited a strong user anthropomorphism (defined as the "ELIZA effect"), with the people attributing understanding, emotion, and intentionality where there was only syntax. This dynamic remains central to contemporary HCI concerns: when systems appear conversationally competent, users may over-ascribe competence and authority, even when underlying mechanisms are narrow and unreliable.

From the late 1980s onward, the dominant paradigm shifted from symbolic specification to a bottom-up approach where models are trained to infer predictive regularities from data called Machine Learning (ML). In ML, intelligence is framed less as explicit rule manipulation and more as the capacity to learn statistical regularities and predictive functions from examples (Mitchell, 1997; Russell & Norvig, 2016). This transition seeded foundational families of approaches—that replaced hand-coded logic with learned representation as the first artificial neural networks or decision trees (Quinlan, 1986; Rumelhart et al., 1986). Pragmatically this paradigm proved more robust in dynamic, noisy environments (e.g., spam classification for emails), yet it also displaced accountability from explicit rules toward parameters, distributions, and training data, raising new questions about how decisions can be inspected, justified, or contested by non-expert stakeholders.

The rise of Deep Learning (DL) intensified this tension. Enabled new architectures, larger datasets, increased computational capacity and advances in training methods, DL networks made it feasible to learn hierarchical representations that outperform earlier approaches in complex domains such as vision, speech and language (Goodfellow et al., 2014; Hinton et al., 2015; Schmidhuber, 2015). Compared to earlier ML pipelines that often relied on human-crafted feature engineering, deep learning systems increasingly co-optimise representation learning and prediction end-to-end. Yet these gains come with a steep cost in legibility: as models increase in depth and parametrization, tracing how inputs influence outputs becomes substantially more difficult. The result is a pragmatic trade: deep models may generalize well, yet their internal logic is harder to summarise in cognitively manageable terms.

In parallel, recent generative architectures have moved AI from prediction and classification toward content production. Generative adversarial networks established a major inflection point in generative modeling (Goodfellow et al., 2014), while transformers architecture enabled scalable language modeling through attention mechanisms (Vaswani et al., 2017). Large Language Models (LLMs), largely enabled by transformer architectures, intensify both the opportunities and the interpretive risks. LLMs are trained via large-scale pretraining and can produce coherent text and adapt to diverse tasks, which expands their applicability in public communication, moderation, summarisation, and participatory workflows. Yet the same linguistic fluency that makes them usable also increases the likelihood of over-trust and "as-if" reasoning by users. Demystifying AI is not an optional educational add-on: it is a condition for responsible adoption. Herrera (Herrera, 2025) captures this well through the idea that advanced technologies can appear "magical" when their mechanisms are not understood; within sociotechnical systems, explainability becomes a way to reduce mystery and align system use with informed judgment.

Against this backdrop, XAI can be understood as a response to the widening gap between model capability and human understandability. The field initially consolidated in a largely technical register. In influential early formulations, interpretability is framed as the "ability to explain or to present in understandable terms to a human" (Doshi-Velez & Kim, 2017), but the operative practices and evaluation regimes often center on technical proxies, model classes and expert-led assessments. Doshi-Velez & Kim (2017) explicitly note the lack of consensus on what interpretability is and how it should be evaluated, proposing taxonomies that range from application-grounded studies with domain experts to functionally grounded evaluations that use formal proxies without humans in the loop (HIP).

This framing also spots a core limit of model-centered explainability. Explanations that satisfy technical criteria (e.g., sparsity, saliency maps, local surrogate models) can still fail as communication artefacts if they do not match the user's task, time constraints, knowledge and context. Doshi-Velez and Kim (2017) explicitly treat the user expertise and time constraints as relevant factors when characterising interpretability needs, and they distinguish human-grounded evaluations that can involve lay participants from functionally grounded approaches that omit human validation. This matter for civic and public-sector contexts: democratic legitimacy depends not only on whether a model can be probed by experts, but also on whether non-expert publics can form warranted judgments about AI-mediated representation and decisions.

In summary, the historical shift from symbolic systems to data-driven deep models explains why explainability has become urgent: contemporary AI often delivers value precisely where its internal reasoning is least tractable to everyday understanding. The same trajectory also marks a shift in XAI from predominantly technical concerns toward human-centered and civic-facing requirements, where explanations must support accountability, contestability, and democratic inclusion—not only debugging or performance assurance.

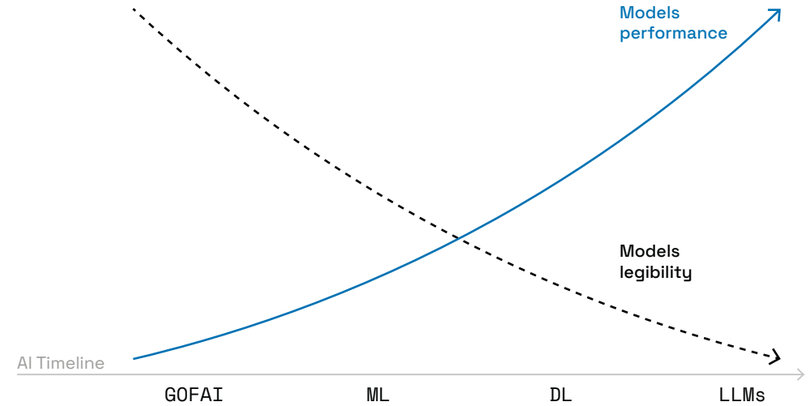


Fig.07 The evolution of AI paradigms: trade-off between performance and opacity

2.2 What Explainability Means in Public and Civic Contexts

AI explainability becomes qualitatively different in civic and public-sector settings: what counts as a meaningful explanation must work for non-expert publics who are affected by, invited to deliberate about, or asked to accept AI-mediated outputs. In these contexts, explainability cannot be reduced to a model property (e.g., feature attributions, rules, saliency maps) or to an "add-on" interface layer appended after deployment. It functions instead as a democratic capability: communicative conditions that let heterogeneous publics understand what an AI-mediated output is doing at a meaningful level, assess its plausibility and limits, and retain the capacity to question, contest, and influence how issues are represented in public reasoning and decision-making.

This reframing clarifies why XAI cannot remain a field for experts only. Historically, interpretability research emerged largely within technical

communities, where the primary audience for explanations was assumed to be model developers, ML engineers, and domain experts (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017). However, as AI systems became embedded in social infrastructures—public administration, welfare, immigration, justice and civic participation—explainability acquired an explicit civic dimension: it became intertwined with public justification, accountability and legitimacy. HCXAI formalized this shift by treating explanation as a sociotechnical and communicative practice shaped by users' compliance and emphasizing the actionability of explanations (Ehsan et al., 2024; Ridley, 2025).

In the current state of XAI research, there are plenty of methods and techniques that try to render Artificial Intelligence models more transparent and understandable, but "more transparency" is not automatically a solution: what matters is whether the resulting account becomes publicly legible and usable for judgment. Buhmann and Fieseler (2023) capture this issue through the notion of "popular comprehensibility", arguing that even if an AI system is explicable among experts, it can remain inscrutable and untraceable for lay publics. They further note that disclosure can paradoxically increase opacity when information volume and complexity overwhelm the citizen. Explainability in civic contexts is less about opening the black box, and more about producing adequate, situated explanations that enable publics to reason with the system's outputs without being forced into expert modes of interpretation.

AI systems are often integrated in domains characterized by asymmetries of power and expertise: welfare, immigration, justice, public consultation and policy evaluation. Here, the normative demand is not simply that systems be accurate, but their outputs can be publicly justified—in the sense that affected people can receive intelligible reasons and oversight actors can assess whether decision processes align with legal, procedural and ethical expectations. This is why policy debates frequently invoke a "right to explanation" (EU GDPR Article 13(2)(f) and 14(2)(g)), even if its legal grounding and scope remain contested. Empirically Van der Veer et al. (2021) show that the absence of a clear guidance on whether AI-based decisions should be accompanied by an explanation is not only a technical issue but a governance gap: explainability is framed as a practical requirement for transparency, accountability, context-sensitivity and attention to societal impact.

From a communication design and HCI perspective, this shifts explainability from a property "inside" the model to a relationship between system outputs, institutional practices, and public sensemaking. In civic contexts,

explainability must therefore support the interpretive work required for democratic participation: evaluating claims, identifying bias or exclusion, and understanding the trade-offs embedded in algorithmic representations.

2.2.1 Plural publics imply plural explanations

A core implication is that there is no single, universal "explanation". Civic AI systems operate across stakeholder groups with different needs and rights: affected individuals, participants in public consultations, caseworkers and moderators, auditors, policymakers, journalists and civil society organizations. Evidence from human-centered XAI research consistently suggests that explanations should be evaluated relative to users' goals, tasks and context rather than against a purely model-centric standard of fidelity and completeness (Rong et al., 2024). This pluralism is also available in deliberative contexts where AI is used to summarize public input, classify arguments, detect themes or structure large-scale discussion. For participants, an explanation may need to clarify why a contribution was clustered, how a summary was generated, or what evidence supports a moderation decision—without requiring exposure to source code or technical internals. For institutional actors, the same system may require process-oriented explanations that enable auditing, documentation and accountability. For publics at large, explainability can mean transparency about the limits of the system, including uncertainty, failure modes and representational bias (Rong et al., 2024).

In democratic settings, the value of explainability is tightly connected to contestability: the capacity of individuals and collectives to challenge AI-generated outputs that shape what becomes visible and actionable in public debate. When AI mediates deliberation at scale, it effectively participates in agenda-settings by filtering, ranking, aggregating and translating citizen contributions into institutional "signals". Under these conditions, an explanation is not only informational; it is a prerequisite for agency, supporting participants' ability to detect misinterpretation, contest outputs and request corrections. This helps specify an important distinction: explainability in civic contexts is less about providing exhaustive causal accounts of model behaviour and more about enabling interpretative access and accountability. The "meaning" of an explanation is therefore pragmatic and situated: it is successful as it supports public reasoning and allows stakeholders to interrogate the system's role in shaping outcomes.

2.2.2 The accuracy-explainability tradeoff

As we mentioned before, a crucial corrective to abstract calls for transparency is that publics do not value explainability uniformly across domains. Van der Veer et al. (2021), using two UK citizens' juries, explicitly test how people trade off explainability and accuracy of the AI system's prediction across different scenarios. They found that the jurors favored accuracy over explainability in healthcare scenarios, whereas in non-healthcare scenarios (recruitment and criminal justice) jurors valued explainability equally or more than accuracy. They also show that justifications for explainability include enabling learning and improvement and strengthening humans' ability to identify and address bias, while accountability concerns emerge differently across contexts. This implies that "explainability requirements" in the public sector should be treated as context and domain specific rather than assumed as a uniform entitlement with identical design solutions.

Because civic explainability targets heterogeneous, often non-expert audiences, it must be designed under realistic cognitive constraints. Cognitive load theory emphasizes that working memory is limited, and comprehension declines when tasks impose excessive processing demands (Sweller, 1988). In practical terms, AI explanations that overload users with feature lists, saliency maps, or technical jargon can undermine understandability, increase reliance on heuristics, and paradoxically reduce meaningful oversight. Human-centered XAI surveys highlight that user studies increasingly measure not only perceived transparency or trust, but also cognitive load and effort as part of explanation evaluation, precisely because explanation interfaces can shift the burden of interpretation onto the user (Rong et al., 2024).

2.3 Structuring the Field of XAI: Global vs Local, Intrinsic vs Post-hoc Approaches

This section introduces a minimal conceptual apparatus to navigate the XAI landscape. Rather than attempting an exhaustive catalogue of methods, it formalizes two widely used structuring axes—scope (global vs. local) and mode (intrinsic/transparent vs. post-hoc)—and clarifies what these distinctions imply for what explanation artefacts can legitimately claim in civic contexts.

In XAI literature, "interpretability" is not a stable object but a bundle of objectives and mechanisms, often invoked as a proxy for trust, fairness, causal insight, or usability. Lipton (2018) frames this as a conceptual problem: interpretability is frequently treated as a monolithic desideratum while being operationalized through incompatible assumptions and heterogeneous proxies. Doshi-Velez and Kim (2017) similarly note the lack of consensus on what interpretability is and how it should be evaluated, advocating shared taxonomies and evidence-based evaluation rather than "you will know it when you see it" reasoning. In civic deliberation this ambiguity is amplified: an artefact that is adequate for practitioners debugging a model may be insufficient—or misleading—when circulated as a public-facing representation for citizens asked to reason about AI-mediated outputs.

2.3.1 Mode of explanation: intrinsic transparency versus post-hoc interpretability

A first recurrent distinction separates transparency (understanding how a model works) from post-hoc interpretability (extracting information after training without exposing the true internal mechanism). Lipton (2018) proposes this bifurcation and further decomposes transparency into levels (e.g., whether a human can simulate the model, whether components are individually interpretable, or whether the training procedure is intelligible). Barredo Arrieta et al. (2020) adopt a closely aligned organization, distinguishing "transparent models" from "post-hoc explainability" and enumerating families of post-hoc techniques (e.g., text, visualization, local explanations, example-based explanations, simplification, feature relevance). Adadi and Berrada (2018) similarly position XAI around "true transparency" versus post-hoc interpretation, emphasizing that post-hoc methods may provide local or global explanations of otherwise opaque models.

For civic deployments, post-hoc methods are often attractive because they can be layered onto already-deployed systems while preserving predictive performance. However, they raise two design-critical risks. First, approximation risk: many post-hoc approaches rely on surrogate models or approximations, which can weaken the faithfulness of the explanation to the original model (Adadi & Berrada, 2018). Second, persuasion risk: explanation artefacts can be compelling and legible while not faithfully describing the underlying decision process, creating "plausible" accounts that may be treated as governance guarantees they cannot deliver (Lipton, 2018). In democratic settings, where explanations can influence belief

formation and perceived legitimacy, this is not only a technical limitation but a communicative and institutional hazard.

2.3.2 Scope of explanation: global, local, and intermediate granularities

The second, orthogonal axis concerns scope: whether an explanation targets the model's behavior as a whole (global) or justifies a particular output (local). Adadi and Berrada (2018) define global interpretability as supporting understanding of the overall logic of the model, whereas local interpretability focuses on explaining a single prediction. Guidotti et al. (2018) formalize the distinction through problem definitions: "black box model explanation" aims to produce a globally comprehensible predictor that mimics the black box, while "black box outcome explanation" aims to provide explanations for individual records. Ribeiro et al. (2016) operationalize local scope through local fidelity—an explanation should reflect model behavior in the neighborhood of a given instance—while stressing that local fidelity does not imply global fidelity.

In civic settings, global explanations align with questions of governance and accountability: what patterns does the system systematically privilege, what evidence does it treat as salient, and under which conditions might it behave unreasonably (Adadi & Berrada, 2018; Guidotti et al., 2018). Yet global scope also surfaces a persistent tension: comprehensibility typically requires simplification, and simplification can conceal conditional behaviors that matter in heterogeneous civic data.

Local explanations, by contrast, justify why the model produced a specific output for a specific case. Ribeiro et al. (2016) frame this in terms of two "trust" problems—trusting a single prediction versus trusting behavior under deployment—and operationalize local explanation in LIME by fitting an interpretable model around an instance, explicitly managing a trade-off between interpretability (bounded complexity) and local fidelity. In civic deliberation, local explanations are relevant because contestation is often triggered by concrete outputs: a flagged message, an extracted "key argument," a predicted stance label, or a ranked list of issues. A local explanation can help convert an opaque output into a debatable proposition by supporting challenges such as "this classification hinges on irrelevant or biased features" or "small changes would reverse the outcome." At the same time, local explanations can become rhetorically powerful while remaining epistemically fragile: a single compelling case-level narrative can be misread as an account of the system as a whole, producing unjustified confidence (Ribeiro et al., 2016).

While "global vs. local" is a useful first cut, it is often too coarse for real deployments. Adadi and Berrada (2018) point to combinations between the extremes, such as modular/global explanations (how parts influence predictions) and group-level explanations (why the model behaves as it does for a cohort). Ribeiro et al. (2016) similarly propose building a broader perspective by selecting representative local explanations (SP-LIME), linking micro-level justifications to macro-level understanding. For civic deliberation, these intermediate granularities are frequently the most actionable: publics rarely need (or can use) a full formal model description, but they often need more than a single-instance rationale to assess whether system behavior is systematically skewed across topics, groups, or contexts.

2.3.3 The joint design space: combining scope and mode

The two axes—scope (global/local) and mode (intrinsic/post-hoc)—define a compact map of the field. They should be treated as design commitments rather than neutral descriptors: they constrain what an explanation can legitimately claim, which stakeholders it can serve, and which failure modes become likely. Crossing the axes yields four canonical regions (intrinsic+global; intrinsic+local; post-hoc+local; post-hoc+global). This is not a complete taxonomy of methods, but a scaffold for making explanatory claims precise and for anticipating civic risks.

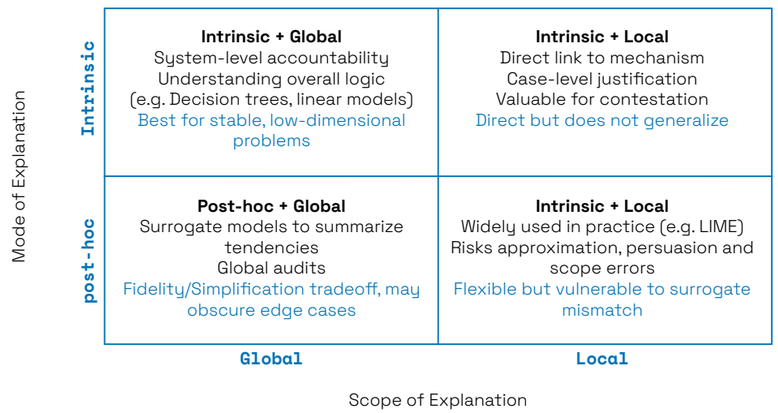


Fig.08 Taxonomy of Explainable AI: scope and mode of explanation matrix

Intrinsic + global approaches support strong system-level accountability narratives, but may fail when the model cannot be simulated or understood

end-to-end under realistic cognitive constraints (Guidotti et al., 2018; Lipton, 2018). Intrinsic + local approaches preserve a direct link to the mechanism while supporting case-level justification, which can be valuable for contestation. Post-hoc + local approaches are widely used in practice (e.g., LIME), but are exposed to scope errors (treating local rationales as global accounts) and to approximation/persuasion risks (Adadi & Berrada, 2018; Lipton, 2018; Ribeiro et al., 2016). Post-hoc + global approaches can provide audits or surrogate models to summarize broad tendencies, but simplification and fidelity limits are especially consequential when outputs are interpreted as public evidence (Guidotti et al., 2018).

2.3.4 Model-agnostic vs model specific

The global/local and intrinsic/post-hoc axes become more analytically powerful when connected to two additional distinctions prominent in the surveys: model-agnostic vs. model specific and the target of explanation. Adadi and Berrada (2018) describe "model-related methods" as either model specific (restricted to a class of algorithms) or model-agnostic (separating prediction from explanation and generalizing over different model architectures). They also note that intrinsic methods are, by definition, model-specific, whereas model-agnostic interpretations are "usually post-hoc" and can be local or global. Model-agnosticism is advantageous in civic infrastructures precisely because it supports institutional flexibility: tools can be applied across vendors, model families and evolving pipelines. Yet the cost is often approximation and weaker guarantees of faithfulness (Adadi & Berrada, 2018; Ribeiro et al., 2016).

2.3.5 From model explanation to pipeline legibility in deliberation

These distinctions are sufficient to establish baseline vocabulary, but they also reveal a structural mismatch with AI-mediated deliberation. Many XAI approaches are organized around explaining decisions affecting identifiable individuals ("Why this prediction?"). In deliberative systems, the outputs of interest are often collective representations produced through multi-stage pipelines (e.g., filtering, clustering, summarization, ranking). The central interpretive question shifts toward "How has discourse been transformed, aggregated, and framed?" Existing taxonomies help specify scope and mode, but leave under-specified how such pipeline-level mediations should be rendered legible when they shape public reasoning. This mismatch also foreshadows the need to render pipeline-level mediations legible when they shape public reasoning.

2.4 Limitations and Risks of Technical Explainability

The previous section mapped the field of XAI by distinguishing scope (global vs. local), mode (intrinsic vs. post-hoc), and coupling to the underlying model (model-specific vs. model-agnostic). That map is analytically useful, but it also clarifies why technical explainability—the production of computational artefacts such as saliency maps, feature attributions, local surrogates, rule lists, or prototypes—cannot be treated as a straightforward solution to the increasing opacity of contemporary models. In civic infrastructures, explanations tend to travel beyond the engineering context in which many methods were originally developed: they are used as reasons for consequential decisions, as evidence of fairness, and as proof of accountability. Under these conditions, the limitations of technical explainability become constitutive risks for democratic deliberation, because they shape what can be understood, challenged, and justified in the first place.

2.4.1 Conceptual underdetermination and the "interpretability" label as an unstable claim

A first limitation concerns the semantic instability of "interpretability" and "explainability." Different approaches encode different definitions of what it means to "open" a black box, and that many proposals specify their interpretability goals only implicitly. This ambiguity is not terminological: it affects what kinds of epistemic and normative claims can legitimately be attached to explanation artefacts. Doshi-Velez and Kim (2017) describe a field where there is "little consensus" on interpretability and, crucially, where evaluation often relies on face-validity logics that approximate "you'll know it when you see it," leaving core questions unresolved (e.g., whether different "interpretable" model classes are comparable, or whether different applications have fundamentally different interpretability needs). Lipton's (2018) critique is especially useful for discipline-building because it explains why "interpretability" is so prone to overreach: the term is frequently used to conflate transparency (how a model operates) with post-hoc accounts (what can be said after the fact about outputs). In civic deployments, this conflation is consequential. When explanation artefacts are circulated as public-facing representations, they can be interpreted as evidence that a system is accountable or fair even when they only provide a narrow, local, or approximate view. The risk is not that explanations are

useless, but that the presence of an explanation becomes a proxy for governance qualities that the artefact, as produced, cannot actually guarantee.

A second, related issue is that interpretability is often invoked as a response to what Doshi-Velez and Kim (2017) call incompleteness: many high-stakes domains cannot fully specify all relevant desiderata (safety, non-discrimination, privacy, or the "right to explanation"), so interpretability is treated as a fallback that enables human verification. Yet incompleteness also implies that explanation artefacts cannot be evaluated solely as technical outputs: their adequacy depends on whether they actually support the forms of scrutiny required by the domain (e.g., identifying discrimination pathways, revealing brittle dependencies, or diagnosing proxy objectives). Where those links are missing, technical explainability risks becoming rhetorically satisfying but institutionally thin.

2.4.2 Fidelity, approximation, and the epistemic gap between explanation artefacts and model behavior

A central technical limitation of many post-hoc approaches is that they rely on approximation. In the surrogate paradigm, an interpretable model is trained or fit to mimic a black-box predictor; the resulting explanation is only as good as the surrogate's ability to capture the black box's decision function. Guidotti et al. (2018) formalize this issue by distinguishing accuracy (agreement with ground-truth labels) from fidelity (agreement with the black box's outputs), noting that fidelity is often operationalized using the same measures as accuracy but with the black box treated as an oracle target. This distinction matters in civic contexts because high-fidelity surrogates can still be socially problematic if the black box itself encodes undesirable behavior; conversely, low-fidelity surrogates can be socially dangerous because they may be interpreted as mechanism-revealing accounts while only weakly tracking the system they purport to explain.

Ribeiro et al.'s (2016) methodological contribution makes this tension explicit in operational terms: complete faithfulness is generally infeasible unless one reproduces the full model, so local explainers must negotiate a fidelity–interpretability trade-off by restricting complexity and focusing on a neighborhood around a specific instance. The civic risk emerges when local explanations—because they are concrete, legible, and persuasive—are used as if they were global accounts of system behavior (a scope error that becomes especially likely in deliberative settings where participants

must reason across cases, categories, and populations). In other words, local explanations can support warranted case-level contestation while simultaneously enabling unwarranted system-level legitimation if their scope constraints are not understood and communicated.

This is also where Lipton's (2018) warning about post-hoc explanation becomes structurally relevant: post-hoc accounts may be practically useful yet fail to disclose how a system works, and they can mislead when they are treated as faithful mechanism-revealing narratives rather than as constrained representations. In civic infrastructures, where explanation artefacts can function as justificatory resources in public debate, the fidelity problem is not merely technical; it is a legitimacy problem, because it reshapes what publics believe they are entitled to know and contest.

2.4.3 Human constraints, representational choices, and the HCI problem of explanation usability

Even where explanation methods are technically well specified, their effectiveness is bounded by human constraints and by the representational decisions embedded in the explanation interface. Guidotti et al. (2018) explicitly foreground time limitation and user expertise as determinants of what counts as interpretable in practice, noting that the literature rarely provides real experiments on the time required to understand an explanation and often substitutes rough complexity proxies (e.g., tree depth, number of rules, number of non-zero weights). Doshi-Velez and Kim (2017) converge on the same point, connecting explanation usability to time constraints and to the "cognitive chunks" with which different users organize information (and therefore to explanation form, quantity, and compositional structure).

These observations have direct implications for civic deliberation. Deliberative settings are characterized by heterogeneous expertise, asymmetric stakes, limited time, and contested interpretive frames. Under such conditions, explanation artefacts can fail in two opposite ways: they can be so reductive that they promote superficial understanding and overconfidence, or so complex that they effectively outsource interpretive labor to non-experts. Adadi and Berrada (2018) discuss this as a cost–benefit tension at the user level: explanations impose cognitive effort and are not automatically beneficial by being available.

The practical issue, from an HCI and communication-design standpoint, is that technical plausibility does not entail usability, and usability is not simply a UI matter but a property of the entire explanatory interaction: what

is selected as salient, how uncertainty is displayed, and how the scope and limits of an artefact are made legible to diverse audiences.

A particularly under-acknowledged source of risk is representation mismatch. Ribeiro et al. (2016) emphasize that many explainers operate on an interpretable representation that differs from the model's actual feature space, separating what the model computes from what humans can meaningfully read (e.g., superpixels instead of raw tensors, bag-of-words indicators instead of embeddings). In civic contexts, where categories are normatively loaded (e.g., "risk," "deservingness," "radicalization," "public sentiment"), representational choices can silently reframe the domain: they decide what counts as a reason, which features are foregrounded as legitimate, and which structural factors become invisible. This makes explanation a mediating artefact rather than a neutral window into computation, and it motivates the shift to HCXAI and communication-design perspectives in the next sections.

2.4.4 Evaluation fragility and institutional misuse: from weak evidence to "transparency theatre"

The limitations above are amplified by a persistent evaluation problem. Doshi-Velez and Kim (2017) distinguish three evaluation regimes—application-grounded (real humans, real tasks), human-grounded (real humans, simplified tasks), and functionally grounded (proxy metrics without humans)—and argue that claims should be calibrated to the evaluation regime used. When evaluation relies primarily on functional proxies (e.g., sparsity, small rule sets, shallow trees), explanation artefacts can appear rigorous while remaining unvalidated for the tasks that matter in civic deployments: enabling contestation, supporting accountability claims, revealing discrimination pathways, or helping publics understand how issues are being represented in AI-mediated debate.

Guidotti et al. (2018) connect these issues to the political economy of opacity and public risk. Their discussion of "black box society" emphasizes that secrecy (industrial confidentiality, legal protections, and other forms of obfuscation) can make discrimination difficult to detect and mitigate, and that delegation to opaque systems without interpretability can create both practical and ethical harms. In public-sector deployments, this creates a structural temptation toward what might be called "transparency theatre": producing explanation artefacts that are institutionally convenient (because they can be shown) but epistemically weak (because they do not materially enable scrutiny). Lipton's (2018) critique of post-hoc

explanation is again relevant here, because the gap between persuasive narratives and faithful accounts can be exploited—intentionally or not—when explanations are treated as public relations objects rather than as instruments of accountability.

Barredo Arrieta et al. (2020) respond by framing explainability as part of a broader responsible-AI agenda, emphasizing that explainability choices must be contextualized by stakeholder needs and by the practical conditions under which explanations are consumed. For civic infrastructures, this implies that technical explainability must be evaluated not only for internal validity (does the method behave as intended?) but also for public validity: does it support the communicative, epistemic, and procedural functions that democratic deliberation demands?

These limitations reveal four structural weaknesses of technical explainability approaches in civic contexts:

- Evaluation regimes often rely on proxy metrics disconnected from real-world deliberative tasks (e.g., contestation, legitimacy assessment, bias diagnosis);
- Explanation fidelity may diverge from persuasive plausibility, enabling accounts that are legible and compelling without being mechanism-revealing;
- Explanations are frequently assessed at the level of individual comprehension rather than collective scrutiny, even though deliberation requires shared interpretive resources and contestable public claims;
- Explanation artefacts can be mobilized institutionally as symbolic transparency ("transparency theatre") rather than as instruments that materially enable accountability.

Consequently, explainability cannot be treated as a detachable technical add-on validated through proxy metrics. In civic infrastructures it must be reconceptualized as a socio-technical capability grounded in users' interpretive practices, representational choices, and the accountability demands that shape what publics can legitimately understand, challenge, and justify.
This motivates the shift toward Human-Centered Explainable AI, which reframes explainability as a socio-technical capability grounded in users,

contexts, and accountability requirements rather than in model properties alone.

2.5 Human-Centered Explainable AI (HCXAI): Principles and Rationale

Human-Centered Explainable AI (HCXAI) begins from a simple but under-addressed premise: explainability is not only a property of models; it is a socio-technical relationship between system, representation, and user, shaped by real constraints (time, attention, literacy, stakes, and institutional power).

Ehsan and Riedl (2020) articulate this shift by distinguishing interpretability (a property that conditions what can be inferred about model behavior) from explanation generation (often post-hoc, aimed at providing accessible information that is useful for users even if it does not fully reveal the model's true internal mechanics). From this perspective, the field is reorganized around the question "Explainable to whom, for what purpose, and under what social conditions?" (Ehsan & Riedl, 2020).

The civic implication is direct: if explanations are treated as self-sufficient technical objects, they can become rituals of transparency—credible-looking outputs that increase acceptance or encourage premature closure without enabling scrutiny, contestation, or learning. Herrera (2025) similarly argues that XAI should support cognitive engagement rather than passive consumption, warning that increasingly capable systems can induce cognitive disengagement and overreliance, undermining critical judgment.

A further adjustment is needed for deliberation. Although HCXAI foregrounds users and contexts, it often still models "the human" as an individual performing discrete tasks (e.g., selecting an option, assessing a recommendation). In AI-mediated deliberation, however, outputs are collective representations that mediate shared discourse. This requires extending HCXAI beyond task-level usability toward plurality preservation, procedural contestability, and legitimacy-sensitive evaluation—i.e., whether publics can interpret, challenge, and justify AI-mediated representations as part of democratic reasoning.

In civic settings, HCXAI responds to a multi-layered problem: (i) explanations are frequently approximations with limited guarantees, (ii) publics and frontline users often lack the technical resources to evaluate those limitations, and (iii) explanation artefacts can become governance objects—

used to justify decisions, allocate responsibility, and shape participation (Ridley, 2025).

2.5.1 Defining HCXAI: from explanations as outputs to explanations as sociotechnical interaction

HCXAI can be defined as a design and research paradigm in which explainability is evaluated primarily by how well it supports human goals, understanding practices, and accountability needs in specific contexts of use—rather than by the existence of an explanation artifact alone (Ehsan & Riedl, 2020; Ridley, 2025).

A distinctive contribution of Ehsan and Riedl (2020) (two of the main figures in this research field) is the insistence that "the human" in HCXAI cannot be treated as a generic, isolated end-user. They advocate a reflective sociotechnical approach that accounts for values, interpersonal dynamics, and the socially situated character of AI systems, supported by methodological strategies in HCI such as value-sensitive design and participatory design, and framed through critical technical practice.

In civic deliberation, this sociotechnical stance matters because AI-mediated outputs (summaries, recommendations, moderation decisions, agenda-setting cues) often operate within a dense web of institutional roles, asymmetries of expertise, and contestable normative assumptions about what counts as a "good" argument, a "representative" summary, or a "relevant" issue.

Herrera (2025) reinforces this reorientation by proposing that definitional cohesion in XAI requires explicitly foregrounding the audience and understanding: explainable AI is not meaningfully "explainable" until the relevant audience can make sense of it for their purposes. This is not an abstract philosophical point: it implies that explanation quality is always conditional on who is interpreting (expert, layperson, regulator, affected community), what they must decide or contest, and how explanation interfaces structure attention and reasoning.

HCXAI treats stakeholders not as a secondary concern but as a primary design variable. Herrera (2025) explicitly describes a multi-level "audience" framing—spanning roles such as developer, designer, owner, user, regulator, and society—and emphasizes that these groups do not share a single notion of "understanding," nor the same requirements for what an explanation should do. This stakeholder plurality aligns with civic

deliberation settings, where explanation may need to support at least four distinct functions simultaneously:

Operational use (e.g., facilitators and participants using AI summaries or topic clustering to navigate complex discussions),
Institutional accountability (e.g., public bodies demonstrating due process, documenting limitations, and enabling review),
Public contestation (e.g., affected communities questioning representation choices, framing assumptions, or data provenance),
Regulatory and audit needs (e.g., evidencing compliance, risk controls, or -transparency obligations).

Ridley (2025) adds an important normative dimension here: those who design and deploy AI systems effectively become "new policymakers," shaping "the rules we live by," which strengthens the argument that HCXAI must be understood as governance-relevant design practice rather than interface embellishment.

2.5.2 Core HCXAI principles for civic-oriented systems

The following principles draw on foundational contributions in XAI and HCXAI literature, interpreted through the lens of civic deliberation as the guiding application domain. They provide conceptual grounding for the thesis; a systematic synthesis of the broader field follows in Chapters 3–4.

1) Audience and task-relative usefulness: Herrera (2025) argues that explainability demands must be framed through the lens of the audience and the practical aim of the explanation, rather than through a generic notion of "transparency." In practice, this means explanations should be designed to support concrete tasks: verifying a summary's representativeness, identifying uncertainty or missing perspectives, checking whether the system is sensitive to irrelevant proxies, or comparing outputs across settings.

2) Interaction over static disclosure: Herrera (2025) criticizes "static, one-size-fits-all explanations," arguing instead for interactive, context-aware and personalized insights that engage users in critical reasoning—helping them "understand, challenge, and refine" AI outputs. Rong et al. (2024) provide empirical support: across user-study evidence, interactivity is among the most consistent factors associated with improved trust, understanding, and usability outcomes.

3) Calibrated trust and reliance: HCXAI does not treat "increasing trust" as a universal success criterion. Scharowski et al. (2023) show why: trust (an attitude) and reliance (a behavior) are distinct, and explanations may increase reliance even when they do not increase attitudinal trust. What matters is trust calibration—whether confidence placed in the system is warranted by its actual capabilities. Herrera (2025) echoes this concern, emphasizing that explainability should preserve human agency and prevent overreliance dynamics.

4) Epistemic humility and limitation awareness: A recurring HCXAI requirement is to communicate limits: what the system can and cannot infer, when explanations are approximations, and how outputs should be used. For civic deliberation, epistemic humility is essential because AI outputs can appear authoritative; interfaces must help users interpret outputs as inputs to reasoning rather than as verdicts.

5) Actionability and contestability: A civic-facing explainability regime is incomplete if explanations do not enable what users can do next. Ridley (2025) highlights actionability and contestability as principles aligned with rights to complaint, recourse, mitigation, and remedial processes; more broadly, they align with democratic expectations of due process and agency. In deliberation infrastructures, this translates to mechanisms to flag misrepresentations, request alternative framings, inspect missing viewpoints, compare summaries, or trigger human review.

6) Comprehensibility as relational semiotic work: Bobek et al. (2025) argue that "comprehensibility" cannot be assumed from the presence of XAI visuals. Drawing on Peirce's semiotic triangle, they conceptualize the explanation artefact as a representamen, the case as the object, and the user's constructed meaning as the interpretant. Different user groups fail in different ways to complete this "semiotic circuit," implying that comprehension is contingent on prior knowledge, trust, and contextual framing. For civic platforms, this supports multiple pathways to comprehension (layered explanations, alternative modalities, scaffolding, and explicit prompts that connect "what you see" to "what it means for your decision").

2.5.3 Evaluation logic: why HCXAI must be empirically grounded in user studies

HCXAI is not only a normative stance; it is a methodological commitment to evaluating explanations in use. Rong et al. (2024) synthesize findings across a large body of user studies and show that explanation effects are mixed and context-dependent: explanations often improve subjective understanding, but their effect on trust and usability is inconsistent, and explanations are not reliably effective at convincing users of model fairness. They also highlight a critical cognitive risk: the illusion of explanatory depth, where users can become overconfident in their understanding of complex systems, especially if understanding is measured only through self-report rather than through tasks that require application of knowledge.

Scharowski et al. (2023) strengthen this point with a measurement argument: if researchers measure only attitudes ("trust") or only behavior ("reliance"), they can draw invalid conclusions about whether explanations support appropriate human use. For civic deliberation—where the target outcome is not compliance but reflective participation—evaluation must therefore include both: (a) whether participants feel informed and empowered, and (b) whether they actually make different, more defensible judgments, detect weaknesses, or contest outputs when appropriate.

2.5.4 Implication for the thesis trajectory: HCXAI as the hinge toward communication design

This section has framed HCXAI as a shift from technical "opening of the black box" to a reflective, sociotechnical, and user-task-oriented approach to explanation. In civic deliberation, the core requirement is not simply legibility but the preservation of agency, contestability, and calibrated reliance under conditions of uncertainty and plural values (Ehsan & Riedl, 2020; Herrera, 2025; Ridley, 2025). The next step is therefore to treat explainability explicitly as a communication and design challenge: how explanation content is selected, narrated, visualized, sequenced, and made interactive becomes a constitutive part of whether publics can understand and legitimately contest AI-mediated representations (Bobek et al., 2025; Rong et al., 2024).

2.6 Explainability as a Communication and Design Challenge

The preceding sections motivate a shift from "explainability as access to model logic" toward explainability as a condition for public sensemaking and accountable use. In civic deliberation, AI systems rarely operate as isolated predictors; they increasingly function as representation engines that transform participatory discourse into artifacts that guide attention and judgment (e.g., summaries, clusters, issue maps, rankings). The critical question is therefore not only whether an explanation can be generated, but whether its communicative form enables diverse publics to understand what has been done, evaluate what can legitimately be inferred, and contest outputs when they misrepresent, oversimplify, or exclude relevant perspectives.

This reframing implies that many explainability failures occur after the model: at the level of representation, framing, sequencing, and interaction. For this reason, explainability becomes a design problem: how to stage complexity, make mediation steps inspectable, and provide verification pathways that non-expert users can actually use under limited time, attention, and heterogeneous literacy.

2.6.1 From model explanation to transformation transparency

When using a communication design approach, explanations should not be treated as a step-by-step reconstruction of algorithm internals, but as an account of what the system did to the deliberation and why that account is meaningful to users in context. In deliberative settings, users may want to know: How did the system transform the material of deliberation into this representation? What was included, excluded, weighted, or merged? What does this representation stand for? Conventional XAI outputs such as feature attributions only partially address these questions, because they remain tied to internal model representations and assume technical literacy that many civic participants do not have.

In this setting, explainability must be re-specified as transformation transparency: the communicative legibility of the mediation pipeline that links inputs (citizen contributions, documents, interactions) to outputs (summaries, clusters, classifications) through identifiable steps (e.g., filtering, deduplication, clustering, summarization, ranking).

Transformation transparency includes, but is not reducible to, provenance. Provenance answers “where did this come from?” by linking an output back to its sources. Transformation transparency additionally answers “what happened in between?” by making visible the intermediate operations that shape meaning. This distinction matters because civic harms often arise not from unknown sources alone, but from hidden transformations—for example, when summarization compresses disagreement into an apparently coherent consensus, or when clustering merges distinct viewpoints under a single label. As Section 2.4 argued, explanation artifacts can be persuasive even when they are only approximate, and local accounts can be misread as global truths.

Operationally, transformation transparency implies that explanation interfaces should do more than provide a rationale; they should support traceable reading—moving from collective representations to evidence, intermediate steps, and boundary conditions of validity.

2.6.2 Scaffolding intelligibility under cognitive constraints

A communication-design approach must also treat interpretability as constrained by attention and cognitive resources. In civic deliberation interfaces, complexity can become a participation barrier when users must interpret dense representations without guidance, compare constructs without framing, or infer the meaning of analytical operations implicitly. This implies that explainability is partly achieved through progressive disclosure and other scaffolds that reduce avoidable interpretation burden while keeping the system open to inspection.

2.6.3 Narrative sequencing and rhetorical accountability

Narrative devices (titles, annotations, contextual framing, staged disclosure) can function as interpretive infrastructure: they can guide sensemaking of complex evidence and AI-mediated outputs. However, narrative is not neutral: sequencing and emphasis can shape salience and interpretation. In civic settings, this requires pairing narrative scaffolds with accountability-oriented design moves (e.g., explicit sourcing, alternative views, uncertainty framing), so that scaffolding does not become a single “official story” but remains open to scrutiny and reinterpretation.

2.6.4 Interactivity as an infrastructure for agency and contestability

If narrative supports entry and comprehension, interaction supports agency. In deliberation, democratic legitimacy depends not only on access to information but on the ability to question and contest it. A baseline interaction grammar is therefore to support movement from overview to details-on-demand, enabling users to traverse from aggregate representations (themes, clusters, summaries) down to evidence traces (source contributions, disagreements, edge cases) and back again. Interactivity becomes a condition for contestability: it allows participants to inspect mediation artifacts and reconstitute meaning rather than merely consume outputs.

2.7 Research Questions

Sections 2.3 to 2.6 progressively reframed explainability from a model-centred concern (“opening the black box”) into a civic requirement: AI-mediated deliberation demands public-facing intelligibility that supports collective sensemaking, scrutiny, and contestation. This shift clarifies why many established XAI and even HCXAI approaches remain only partially adequate for democratic contexts. Technical explanations often succeed at producing an account of model behaviour (Lipton, 2018; Ribeiro et al., 2016), yet civic deliberation requires something broader: the ability for heterogeneous publics to understand how discourse is transformed into representations, to interpret those representations appropriately, and to challenge them when they distort, exclude, or overclaim. In other words, the core problem is no longer only the opacity of models, but the opacity of mediation—how computational transformations reorganize public discourse into actionable artifacts.

The research questions below are formulated to (i) be answerable through the SLR as an evidence-mapping exercise across the three pillars, (ii) motivate a communication-design framework for civic explainability, and (iii) support the later development and discussion of a concrete artifact as grounded instantiations of the framework.

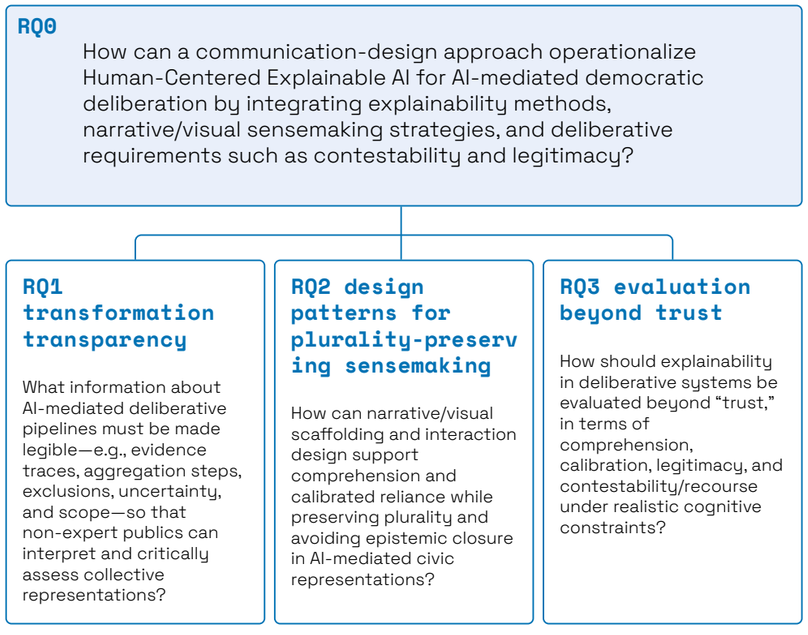
RQ0
How can a communication-design approach operationalize Human-Centered Explainable AI for AI-mediated democratic deliberation by integrating explainability methods, narrative/visual sensemaking strategies, and deliberative requirements such as contestability and legitimacy?

RQ1 (transformation transparency)
What information about AI-mediated deliberative pipelines must be made legible—e.g., evidence traces, aggregation steps, exclusions, uncertainty, and scope—so that non-expert publics can interpret and critically assess collective representations?

RQ2 (design patterns for plurality-preserving sensemaking)
How can narrative/visual scaffolding and interaction design support comprehension and calibrated reliance while preserving plurality and avoiding epistemic closure in AI-mediated civic representations?

RQ3 (evaluation beyond trust)
How should explainability in deliberative systems be evaluated beyond "trust," in terms of comprehension, calibration, legitimacy, and contestability/recourse under realistic cognitive constraints?

Fig.09 Reaserch questions structure



Chapter 2 has progressively reframed explainability from a predominantly model-centred concern into a civic requirement. Technical XAI approaches, while valuable, often provide partial and easily misread accounts, that HCXAI usefully shifts attention toward users and contexts yet remains under-specified for multi-stakeholder democratic settings, and that the decisive challenges become communicative and design-based—how discourse is transformed into representations, how value-laden parameter choices are rendered legible, and how narrative coherence is achieved without epistemic closure. These research questions motivate the need for an integrative evidence base that spans explainability research, narrative/visual communication, and deliberative theory. Chapter 3 therefore introduces the systematic literature review that operationalizes this three-pillar framing, specifying the search strategy, inclusion criteria, and coding approach used to map current knowledge, surface cross-field blind spots, and derive design-oriented requirements that will ground the framework and artifact development in the subsequent chapters.

3.0

RESEARCH THROUGH DESIGN METHODOLOGY. FROM SYSTEMATIC REVIEW TO REAL-LIFE EXPERIMENTATION

This chapter describes the methodology adopted to conduct a systematic literature review that informs the thesis contribution at the intersection of three domains: (i) Explainable AI and Human-Centered XAI (XAI/HCXAI), (ii) data visualization and data storytelling as sensemaking practices, and (iii) democratic deliberation and AI-mediated civic participation. Chapter 2 argued that explainability in civic contexts cannot be reduced to a technical property of models: it must operate as a communicative infrastructure that enables heterogeneous publics to interpret AI-mediated representations, assess their limits, and contest them when necessary. To translate this conceptual framing into a grounded research program, the thesis requires an evidence base that spans and connects these otherwise fragmented literatures.

A systematic review is particularly appropriate for this purpose because relevant contributions are distributed across disciplines with different terminologies, publication venues, and evaluative traditions. Rather than aiming to benchmark methods or produce an exhaustive catalog of techniques, the review is designed to (i) map how explainability is framed across the three pillars, (ii) identify recurring design-relevant concepts and failure modes (e.g., transformation transparency, value-laden parameterization, narrative risks), and (iii) derive synthesis outputs that can inform a communication-design framework and support the justification of later design artifacts. The review functions as a bridge between conceptual analysis (Chapter 2), empirical synthesis (Chapter 4), and framework- and artifact-oriented contributions (Chapters 5–8).

The chapter is organized as follows. Section 3.1 describes the data sources and query strategy used to retrieve candidate studies across Scopus and Web of Science. Section 3.2 specifies the inclusion and exclusion criteria, reflecting the thesis' cross-pillar focus and emphasis on human-centered interpretation and civic relevance. Section 3.3 outlines the screening and selection process leading to the final corpus, while also clarifying how foundational and snowballing references complement the core set of systematically retrieved papers. Section 3.4 details the data extraction and coding scheme used to translate papers into analyzable evidence, and Section 3.5 explains the thematic synthesis approach used to generate higher-level findings. Section 3.6 discusses limitations and reflexive considerations that contextualize the scope of the review. Section 3.7 explain how the collaboration with the ORBIS project was structured and section 3.8 how the thesign process was organized.

3.1 Search Queries and Data Sources

This systematic review was designed to map and integrate three research pillars that jointly define the thesis problem space: (1) Explainable AI and Human-Centered XAI (XAI/HCXAI), (2) data visualization and data storytelling as sensemaking infrastructures, and (3) democratic deliberation and AI-mediated civic participation. Because relevant contributions are distributed across fields that use partially incompatible vocabularies, the review combined an initial scoping step with a structured database search.

Scoping phase

The research began with an exploratory scoping activity aimed at identifying foundational references and stabilizing terminology across the three pillars. This step was conducted using Google Scholar by querying each topic separately and ranking the top results by citation count. Since Google Scholar does not directly support ordering by citations, the citation ranking was obtained through a programmatic extraction process. The purpose of this step was not to build the systematic review corpus, but to establish a shared conceptual baseline and to inform the construction of the subsequent database queries.

Systematic database search

The systematic search was conducted on Scopus and Web of Science between 03/12/2025 and 04/12/2025, following a structured query strategy designed to capture different intersections of the three pillars. Searches were performed using TIT-ABS-KEY() fields in Scopus and the TS=() field in Web of Science, ensuring that retrieved records were topically anchored in titles, abstracts, and keywords rather than in full-text indexing. Seven queries (Q1–Q7) were defined to cover: (i) visualization for explaining AI, (ii) storytelling and narrative visualization in AI-related contexts, (iii) explicit links between storytelling and XAI, (iv) AI and deliberative democracy, (v) a broader combined query spanning visualization, storytelling, and AI, (vi) visualization and deliberation, and (vii) explicit HCXAI terminology.

Tab. 01 Query strings for the systematic literature review

Query	Conceptual focus	Query string (as executed)
Q1	Data visualization for explaining AI	(“explainable AI” OR “explainable artificial intelligence” OR “XAI” OR “model interpretability”) AND (“data visualization” OR “data visualisation” OR “visual analytics”)
Q2	Storytelling + data visualization + AI	(“data visualization” OR “data visualisation”) AND (“data storytelling” OR “narrative visualization” OR “narrative visualisation” OR “scrollytelling”) AND (“artificial intelligence” OR “AI” OR “algorithmic decision” OR “large language model” OR “LLM”)
Q3	Storytelling + XAI	(“data storytelling” OR “narrative visualization” OR “narrative visualisation” OR “data stories” OR “scrollytelling”) AND (“explainable AI” OR “explainable artificial intelligence” OR “XAI” OR “model interpretability”)
Q4	AI and deliberation	(“artificial intelligence” OR “AI” OR “algorithmic decision” OR “large language model” OR “LLM”) AND (“deliberative democracy” OR “democratic deliberation” OR “public deliberation” OR “online deliberation”)
Q5	Broad intersection: visualization + storytelling + AI	(Scopus equivalent / WoS TS=) (“data visualization” OR “data visualisation”) AND (“data storytelling” OR “narrative visualization” OR “narrative visualisation” OR “scrollytelling”) AND (“artificial intelligence” OR “AI” OR “machine learning” OR “algorithmic” OR “model” OR “complex systems”)
Q6	Data visualization and deliberation	(“data visualization” OR “data visualisation” OR “visual analytics”) AND (“deliberative democracy” OR “democratic deliberation” OR “public deliberation” OR “online deliberation”)
Q7	HCXAI	(“human-centered explainable AI” OR “HCXAI”)

The combined searches produced 309 initial hits across databases. After removing records that were not research papers (e.g., non-article items or metadata artifacts), the dataset consisted of 303 papers. Records were exported from Scopus and Web of Science in BibTeX format, imported into Zotero (organized by database and query), and then exported as CSV files for consolidation and screening.

3.2 Inclusion and Exclusion Criteria

To ensure coherence across the three-pillar research scope and to maintain a consistent human-centered orientation, candidate studies retrieved through database queries were screened using a set of inclusion and exclusion criteria. The criteria were designed to filter out purely technical contributions and single-domain work, while retaining studies able to inform

the thesis’ central concern: how AI-mediated representations can be made intelligible, usable, and contestable for heterogeneous publics in civic and socially consequential settings.

Inclusion Criteria

Studies were included if they met all of the following conditions:

- Peer-reviewed publication
- Articles published in peer-reviewed journals or conference proceedings and available as final versions (no preprints, drafts, or extended abstracts).
- Written in English

Relevance to the research domain

Studies substantively engaging with at least two of the thematic pillars of this research:

- Explainable Artificial Intelligence (XAI) or Human-Centered XAI (HCXAI)
- Data Visualization (or Visual Analytics) and Data Storytelling (or DataNarrative Visualization)
- Democratic deliberation, public participation, or civic decision-making

Contribution to human-centered interpretation or sensemaking

Studies providing—to various extents—insights into how people interpret, understand, or act upon AI-generated or data-driven information. This includes:

- Mechanisms of transparency, explainability, or interpretability
- Visual or narrative scaffolds for comprehension
- Cognitive, social, or communicative dimensions of AI mediation
- Tools, models, frameworks, or systems for enhancing user agency, understanding, or trust

Applicability to public, civic, or socially relevant contexts
Studies either directly addressing civic/democratic settings or providing transferable human-centered design insights relevant to such contexts.

Exclusion Criteria

Studies were excluded if they met any of the following:

- Purely technical or algorithmic work
  Studies focusing exclusively on model architecture, optimization, performance benchmarking, or computational techniques without addressing human understanding, interaction, or interpretability.

- Single-domain focus
  Papers addressing only one thematic without being related or transferable to the other pillars were excluded.

- Commercial or private-sector focus with limited relevance for public sector/civic domain
  Studies centred narrowly on business intelligence, marketing, industrial optimization, or proprietary systems with no implications for civic participation, public sensemaking, or democratic contexts.

- Irrelevant context or lack of user-oriented contribution
  Papers that do not provide insights into human cognition, communication, interaction, transparency, or participatory dimensions of AI and data systems.

- Non-peer-reviewed documents
  Documents including opinion pieces, blog posts, white papers, magazine articles, or grey literature unless accompanied by peer-reviewed evidence.

Clarification: core corpus and supporting foundational set

While the systematic screening process prioritized cross-pillar engagement (≥2 pillars) to construct the core review corpus, a limited number works were retained as a supporting foundational set when they contributed essential conceptual anchors for interpreting findings across the three pillars. These supporting papers are analytically distinguished from the core corpus and are used for framing and cross-validation rather than for characterizing the core evidence base.

3.3 Screening and Selection Process

The screening and selection process was conducted in successive stages to progressively refine the initial retrieval into a final review corpus that matched the cross-pillar scope and human-centered focus of this thesis. Searches across Scopus and Web of Science produced 309 initial hits. After removing records that were not research papers (e.g., non-article items and database artifacts), the dataset consisted of 303 papers. These records were imported into Zotero and then consolidated into a single spreadsheet for screening. Duplicate entries were removed using a composite matching procedure based on title, author(s), and publication year, resulting in a deduplicated dataset of 277 papers.

Screening was then performed in a tiered manner, moving from broad topical relevance to more precise assessment of contribution and fit with the inclusion criteria. First, a title screening removed clearly out-of-scope records, yielding 132 papers. Second, an abstract screening further filtered the dataset for cross-pillar relevance, human-centered interpretive contribution, and civic or socially consequential applicability, yielding 80 papers. Third, as an intermediate step before full-text reading, introductions and conclusions were screened to confirm that papers made a substantive contribution aligned with the review scope; 39 papers were retained for full-text analysis. Finally, the full-text screening resulted in 23 papers that formed the core corpus for the systematic review.

In addition to the core corpus, five papers were added through a complementary process of snowballing and foundational anchoring, resulting in a final review set of 28 papers. These additional papers are treated as a supporting set (Section 3.2), included to strengthen conceptual grounding and cross-validation of findings, while remaining analytically distinguishable from the systematically retrieved core corpus.

3.4 Data Extraction and Coding Scheme

To translate the final set of studies into analyzable evidence, each paper was subjected to a structured data extraction process guided by an evolving codebook. The aim of the coding strategy was not only to record bibliographic and methodological descriptors, but to surface design-relevant constructs for civic explainability—i.e., how studies conceptualize interpretability as a communicative problem, how they propose representational and interactional solutions, how narrative visualization

can be implemented in the explanation process, and how they evaluate explanation effectiveness beyond narrow “trust” measures.

3.4.1 Data extraction procedure

For each paper, a standardized extraction template was used to capture:

- Descriptive metadata (e.g., publication year, venue, discipline signal, and study type);
- Domain and context (e.g., civic/deliberative relevance, type of system or decision context);
  Explanation target and medium (e.g., what is being explained, to whom, and through which representational forms such as text, visualizations, narratives, interactive interfaces);
- Human-centered focus (e.g., assumptions about users and literacy, sensemaking goals, interpretive risks);
- Evaluation logic (e.g., outcomes assessed, methods used, and whether evaluation addresses comprehension, calibration, legitimacy, contestability/recourse);
- Design implications explicitly stated by authors and those inferred through recurring patterns across papers.

3.4.2 Hybrid coding logic and codebook evolution

Coding followed a hybrid deductive–inductive strategy. An initial set of high-level codes was defined to reflect the core analytical commitments of the thesis and to ensure comparability across studies. During full-text reading, additional open codes were created whenever recurring concepts, mechanisms, or failure modes emerged that were not adequately captured by the initial schema. When an open code recurred across multiple papers and proved analytically useful for distinguishing design implications, it was promoted into the codebook and integrated into subsequent extraction. This iterative process is documented through successive codebook versions (v1.0 to v1.5), which progressively increased conceptual specificity and improved the fit between the coding scheme and the interdisciplinary evidence base. The codebook thus served as both (i) an extraction instrument for consistent analysis and (ii) a trace of conceptual refinement as the review progressed.

3.4.3 High-level analytical dimensions

The initial code set was organized around the following analytical dimensions, which reflect the thesis’ framing of explainability as a communication and design challenge:

1) Sensemaking and interpretive resources: users’ mental models, literacy assumptions, explanation uptake, and common misinterpretations.
2) Representation and rhetoric: framing strategies, narrative devices, metaphors, visualization techniques, and evidentiality cues used to render AI-mediated outputs meaningful.
3) Interaction and control: progressive disclosure, user agency in exploring explanations, personalization, alternatives, and comparative views (“why/why-not/what-if” styles of inquiry).
4) Uncertainty, limits, and epistemic humility: how uncertainty and limitations are communicated (or obscured), including confidence cues, disagreement, scope boundaries, and contestable claims.
5) Deliberative mediation artifacts: mechanisms through which explanations support collective discussion (e.g., shared references, coordination, argumentation scaffolds).
6) Accessibility and inclusion: plain language strategies, multilingual considerations, cognitive accessibility, and design for heterogeneous publics.
7) Breakdowns and risks as design problems: overtrust, persuasion/compliance drift, cognitive overload, strategic manipulation through framing, and other failure modes.

These dimensions provided a stable scaffold for coding across the corpus while remaining compatible with inductive refinement through open coding.

3.4.4 AI-assisted extraction and consistency across access constraints

Because the corpus included both open-access and restricted-access publications, the practical coding workflow combined manual interpretation with selective tool support. Open-access papers were open to AI-assisted

extraction to improve consistency and reduce repetitive transcription work, while restricted-access papers were coded manually following the same extraction structure established through the codebook. Across both cases, the analytic goal remained consistent: to code each paper against the shared schema and to capture additional open codes when novel, design-relevant concepts emerged.

3.4.5 Output of the coding process

The primary output of this process is a coded dataset that supports two forms of synthesis: (i) a thematic mapping of how explainability, storytelling/visualization, and deliberative requirements are currently connected (Chapter 4), and (ii) the derivation of design-oriented requirements and principles for civic explainability (Chapter 5), which later inform the development and justification of research-through-design artifacts (Chapters 6–9).

3.5 Thematic Synthesis and Analytical Approach

Following screening and structured extraction/coding, the final corpus was synthesized through a framework-informed thematic synthesis aimed at producing cross-disciplinary, design-relevant insights. The purpose of the synthesis was to (i) integrate evidence across the three research pillars (XAI/HCXAI; data storytelling/visualization; democratic deliberation), (ii) identify recurring mechanisms, tensions, and failure modes in how explainability is conceptualized and operationalized for human interpretation, and (iii) translate these patterns into inputs that could be used to structure a design framework for the artefacts.

3.5.1 Unit of analysis and synthesis inputs

The unit of analysis was the individual study. Each paper was treated as a case described through a standardized extraction template and coded using (a) the stable dimensions of the codebook and (b) inductively generated open codes captured during reading. Synthesis was performed on the basis of an extraction matrix that enabled cross-case comparison across: problem framing, intervention and explanation target, evidence base/evaluation approach, key findings and limitations, transferability to civic contexts, and coded analytical dimensions.

3.5.2 Framework-informed thematic synthesis with inductive refinement

The synthesis combined two complementary logics:

Framework guidance (RQ-aligned): candidate thematic clusters were guided by the research questions defined in Section 2.7, with particular attention to (i) transformation transparency (RQ1), (ii) design patterns for plurality-preserving sensemaking (RQ2), and (iii) evaluation beyond trust (RQ3). This ensured that the synthesis remained aligned with the thesis' communication-design contribution rather than reproducing field-specific taxonomies.

Inductive refinement (open-code driven): within and across these clusters, recurring open codes were used to refine boundaries, identify sub-themes, and capture emergent concepts that were not fully anticipated by the initial schema. Open codes functioned as "theme seeds" and were used to articulate specific mechanisms (e.g., representational strategies, interaction patterns, epistemic cues, or breakdowns) that recur across different disciplinary framings.

3.5.3 Theme construction procedure

Theme construction proceeded iteratively through four steps:

1) Cross-case comparison: extracted evidence was compared across papers to identify recurring communicative problems, representational strategies, interaction patterns, and evaluation targets.

2) Clustering into candidate themes: related codes and open-code concepts were grouped into candidate clusters. Clusters were organized around the communicative and design functions that explanations were expected to serve (e.g., making mediation legible, scaffolding interpretation, enabling contestation, supporting calibrated reliance).

3) Refinement and boundary-setting: clusters were merged or split when they were either too broad (masking important distinctions) or too narrow (capturing paper-specific peculiarities). Boundaries were defined by specifying what kinds of evidence and mechanisms counted as instances of a theme and what did not.

4) Traceability check: each theme was retained only when supported by multiple papers and by explicit extracted evidence (claims, methods, findings, or design implications). Contradictory or task-contingent evidence was preserved as part of theme definition rather than treated as noise.

3.5.4 Meta-theme organization (MT1–MT4)

For synthesis management and to support a readable findings structure, papers and extracted evidence were organized into four meta-themes. These meta-themes are not treated as exclusive categories; several papers contribute to more than one meta-theme, reflecting the interdisciplinary nature of the corpus and the fact that communicative, interactional, and evaluative aspects of explainability often co-occur in the same contribution.

- **MT1** - Transformation transparency of deliberative mediation: evidence concerning what information about AI-mediated pipelines must be made legible (e.g., aggregation steps, exclusions, provenance, scope, uncertainty) to support interpretation and critical assessment of collective representations (RQ1).

- **MT2** - Narrative and visual scaffolding for plurality-preserving sensemaking: evidence on representational and narrative strategies that support comprehension and calibrated reliance while preserving plurality and reducing risks of epistemic closure (RQ2).

- **MT3** - Interaction, agency, and contestability mechanisms: evidence on interaction patterns and control structures that enable users to explore, interrogate, and contest AI-mediated representations, including recourse and deliberative agency (RQ2–RQ3).

- **MT4** - Evaluating democratic adequacy beyond trust: evidence on evaluation targets and methods that go beyond generic "trust," focusing on comprehension, calibration, legitimacy, and contestability/recourse under realistic cognitive constraints (RQ3).

Across these meta-themes, the integrative research question (RQ0) functions as an overarching synthesis lens: it motivates the combined consideration of representational, interactional, and evaluative dimensions as a single communication-design problem, rather than treating them as independent concerns.

Tab. 02 Codes mapped to the meta-themes defined from the SLR

Meta-theme	Codebook - primary codes inside the meta-theme	supporting
MT1: Transformation transparency of deliberative mediation (RQ1)	Cross-modal evidential anchoring; Uncertainty/limits communication; Salience & visibility orchestration (ranking/routing/clustering); Normative compression / proxy operationalization; Epistemic alignment / domain legibility.	Evidential trust / credibility tension; Explanation selection heuristics & bias; Contrastive explanation; Stage-based feature mapping; Facilitation–control boundary; Breakdowns & failure modes.
MT2: Narrative and visual scaffolding for plurality-preserving sensemaking (RQ2)	Representation & rhetoric; Narrative sequencing & staging; Sensemaking scaffolds; Accessibility & inclusion; Semiotic completion / semiotic circuit; Autotelic explanation reception.	Componentisation / modular narrative grammar; Genre composability / containment relations; Explanation selection heuristics & bias; Contrastive explanation / fact–foil framing; Cross-modal evidential anchoring; Breakdowns & failure modes.
MT3: Interaction, agency, and contestability mechanisms (RQ2–RQ3)	Interaction & control; Authoring agency / narrative reconfiguration; Defensive design / epistemic vigilance affordances; Cross-modal evidential anchoring (contestability infrastructure); Facilitation–control boundary.	Model-as-boundary-object for deliberation; Stage-based feature mapping (pipeline fit); Contrastive explanation; Explanation as learning infrastructure; Sensemaking scaffolds; Salience & visibility orchestration; Breakdowns & failure modes.
MT4: Evaluating democratic adequacy beyond trust (RQ3)	Evaluation validity / measurement vigilance; Task-contingent effects; Evidential trust / credibility tension; Breakdowns & failure modes.	Uncertainty/limits communication (as evaluable communicative condition); Epistemic alignment / domain legibility; Accessibility & inclusion; Autotelic explanation reception; Explanation selection heuristics & bias; Defensive design / epistemic vigilance affordances.

3.6 Review Limits and Reflexive Considerations

This review was designed as a structured, interdisciplinary evidence base to support a communication-design contribution on explainability for AI-mediated democratic deliberation. As such, it prioritizes cross-field integration and design-relevant synthesis over exhaustive coverage of any single discipline. The following considerations delimit the scope of the review and contextualize the interpretation of its findings.

First, the systematic search was restricted to Scopus and Web of Science in a specific time frame, and knowing the fast growing rate of the Artificial Intelligence field, new relevant contributions will be published or are already available without being included in the literature review. While these databases provide broad coverage of peer-reviewed literature and support reproducible querying, their indexing practices may under-represent relevant contributions published in venues with weaker database coverage, including certain design-oriented outlets and emerging interdisciplinary forums. In addition, terminology varies significantly across the three pillars; despite the use of multiple queries to capture different intersections, some relevant work may not have been retrieved due to vocabulary differences or the absence of standard keywords.

Second, the review corpus reflects a cross-pillar inclusion strategy: the core dataset prioritized studies that substantively engaged with at least two pillars (XAI/HCXAI; data storytelling/visualization; democratic deliberation). This criterion strengthens alignment with the thesis scope but can exclude papers that are highly influential within a single pillar. To mitigate this phenomenon, single-pillar papers with a very strong contribution and human-centered and civic relevance were kept in the analysis, and in addition some single-pillar foundational papers entered the review through a snowballing process. This separation preserves the interpretive value of conceptual anchors while maintaining transparency about what constitutes the systematic evidence base.

Third, methodological heterogeneity across included studies constrains direct comparability. The final set contains empirical user studies, conceptual frameworks, system papers, and surveys, each with distinct assumptions about users, tasks, and evaluation. The synthesis therefore emphasizes patterns in communicative functions, representational strategies, interaction affordances, and evaluation targets rather than attempting to aggregate results through quantitative comparison.

Finally, the review is conducted within the broader logic of a design-led research trajectory: its goal is not only to summarize existing work but to derive actionable insights for framework development and artifact-oriented exploration. Consequently, the synthesis is oriented toward identifying design tensions, failure modes, and operationalizable requirements—particularly around transformation transparency, plurality-preserving sensemaking, contestability, and evaluation beyond trust. These limitations do not undermine the value of the review; rather, they define the boundaries within which the findings should be read and the context in which subsequent chapters develop and test design-oriented responses.

3.7 Study setting: Research Through Design in ORBIS AI-Supported Deliberation

This thesis was conducted within the Horizon Europe project ORBIS, coordinated by Politecnico di Milano, involving a multi-partner consortium (13 participants and 2 partner organisations).

3.7.1 The ORBIS project

ORBIS addresses a recurring limitation of contemporary digital participation: deliberative inputs are often fragmented and difficult to interpret at scale, weakening the connection between citizens' contributions and institutional decision-making. In response, ORBIS develops a socio-technical approach to support a more inclusive, transparent, and trustworthy form of digitally enhanced deliberative democracy, with particular attention to scaling participation while preserving deliberative quality.

A core output of ORBIS is an AI toolkit designed to process large volumes of deliberative input and produce structured, intelligible representations of public discourse. The toolkit comprises four components: Argument Mining (AM) (to extract and structure argumentative content), Feedback Aggregator (FA) (to cluster, enrich, and summarise contributions), Explanation Generator (EG) (to provide human-readable explanations of AI operations), and Policy Recommendation (PR) (to transform deliberative outputs into actionable policy suggestions). These components are integrated into three deliberative platforms used within the project: BCause (structured pro/con discussions), PolisOrbis (large-scale participation based on the Pol.is engine), and Democratic Reflection (real-time audience feedback for live or recorded events).

Methodologically, ORBIS is characterised by a structured co-creation and co-design approach anchored in six real-world pilots. Pilots operate as experimentation settings in which tools are iteratively specified, refined, and validated with civic actors and organisers. The project's co-creation methodology is organised into three sequential phases (plus a parallel synthesis layer), progressing from needs mapping to iterative co-design and contextual integration of AI-enhanced tools. ORBIS also adopts human-in-the-loop oversight, where AI-generated content is subject to facilitator/moderator validation, supporting accountability in deployment.

Within this thesis, ORBIS provided the opportunity to conduct a real-world experimentation in which findings from the literature review and the derived design framework could be tested and aligned with project constraints and requirements.

3.7.2 KG in BCause platform

Inside ORBIS, my practical work was situated in the Deliberation Visualisation and Analytics stream and focused on the Knowledge Graph (KG) as a visual analytics tool applied to BCause deliberative discussions.

BCause is one of the deliberation platforms integrated in ORBIS, designed to host and manage large-scale, text-based deliberations in a structured way. Participants submit free-text contributions (e.g., opinions, arguments, reactions) that the platform organises into an argumentative discussion space. In ORBIS, BCause is enhanced through the integration of the project's AI toolkit (Argumentation Mining, Feedback Aggregator, Explanation Generator, Policy Recommendation), which processes deliberation data to support moderation, synthesis, and analysis. In particular, ORBIS develops mechanisms to (i) classify and structure content into argumentative components, (ii) generate summaries and keyphrases, (iii) support clustering and knowledge graph enrichment, and (iv) produce explanations meant to keep AI outputs intelligible and accountable within the user workflow.

At the interface level, ORBIS extends BCause with a reporting and analytics layer primarily used by discussion administrators (moderators). Moderators can generate reports at different moments during an ongoing debate, producing an archive of time-stamped "snapshots" that can be revisited and compared. Each report is presented as an interactive dashboard composed of multiple widgets that combine raw discussion data with outputs from the ORBIS toolkit. The dashboard is configurable (widgets can be rearranged and resized) and supports exporting individual widgets

(e.g., PNG) or the full dashboard (e.g., PDF), which is relevant for facilitation, documentation, and sharing.

In this ecosystem, the Knowledge Graph (KG) is one of the visual analytics instruments used to make complex deliberation more legible and navigable, complementing more linear representations (e.g., chronological reading or platform-native views such as argument-oriented structures). The KG translates the discussion into a relational representation that explicitly encodes how key deliberative elements connect, relations among positions, arguments, recurring themes/concepts, and contributions. This matters because, as the volume of participation grows, the density of interconnections quickly exceeds what users can reconstruct by reading contributions sequentially. A graph-based representation is therefore used to surface structure that would otherwise remain implicit in the discourse.

The KG is important in ORBIS for three methodological reasons:

1) Sensemaking for participants (orientation without exhaustive reading).
   By turning the debate into a connected structure, the KG helps users quickly locate where agreement, disagreement, and uncertainty concentrate and how different viewpoints relate at a thematic level—supporting exploratory navigation rather than requiring full linear consumption of content.

2) Diagnostic support for facilitators and analysts (debate monitoring).
   The KG enables facilitation-oriented questions such as: which positions are central in the debate, where polarisation emerges, what arguments connect otherwise separated viewpoints, and which themes cut across multiple positions. This supports moderation and synthesis activities, especially in large-scale settings.

3) Human-centred explainability and traceability.
   Within ORBIS, the KG operationalises a human-centred approach where AI is used to surface patterns and structure, while interpretation remains with human actors. Crucially, the KG is intended to preserve traceability between individual contributions and higher-level representations (e.g., themes or clusters), so that collective views remain inspectable and contestable rather than "black-box" summaries.

Overall, in the ORBIS implementation, BCause provides the structured deliberation environment and reporting workflow, while the KG provides a complementary structural map of the debate, supporting sensemaking, facilitation, and explainable navigation of complex civic discourse at scale.

3.7.3 Data Sources and Materials

This study was grounded in two complementary inputs: (i) a filtered subset of ORBIS requirements (originally defined in D2.2 and later referenced in ORBIS technical documentation), used to frame what the Knowledge Graph (KG) should enable from a user and analysis perspective; and (ii) the data exports and APIs made available within the ORBIS infrastructure, used as concrete design materials for the KG development.

3.7.3.1 ORBIS requirements from D2.2

A set of ORBIS requirements originating from ORBIS Deliverable D2.2 was analysed and screened to identify those relevant for a user-facing Knowledge Graph for deliberative discussions. The filtering was conducted by reviewing each requirement and marking its relevance for this study (Yes/No, with “strong” relevance where the requirement directly targets graph exploration or discussion analysis capabilities).

The original requirements form the ORBIS documentation are clustered in these categories:

1) User Interaction and Engagement (UIE): Requirements that involve the overall user experience and engagement with the system

2) Discussion Analysis and Visualization (DAV): Requirements regarding the examination and representation of textual discussions

3) Moderation and Assistance (MA): Requirements focusing on the facilitation of effective collaborative discussion

4) Reports and Summarization (RS): Requirements centered around the generation of informative summary reports

5) Collaborative Features (CF): Requirements fostering collaboration and interaction among users

6) Multi-Phase Deliberations (MPD): Requirements about the complex decision-making process that is carried in multiple stages

Tab. 03 Selection of requirement from the ORBIS project that are relevant for this study

ORBIS req. ID	Requirement	Relevance for this study
UIE.1	The system stores user profiles	Yes
UIE.2	Notification mechanism for event updates	No
UIE.3	Personal space (idea resource pad) for each user.	Yes
UIE.4	User-friendly interface for exploring network graphs of ideas.	Yes (strong)
UIE.5	Real-time transcription of live discussion	No
UIE.6	User recommendation according to user profile	No
DAV.1	Cluster discussion data into main viewpoints/arguments.	Yes (strong)
DAV.2	Summarization of main arguments of discussion.	Yes
DAV.3	Identification of key influential people/actors.	Yes (strong)
DAV.4	Comparison of argument graphs and detection of overlaps and disjoints.	Yes (strong)
DAV.5	Real-time analysis and classification of key elements of discussion.	No
DAV.6	User-friendly interactive interface for exploring viewpoints.	Yes (strong)
MA.1	Expert collaboration space to assist policy formulation.	No
MA.2	Minutes creation and structured discussion interface.	Yes (partially)
MA.3	Feedback mechanisms (e.g., voting, thumbs up/down, etc.).	No
MA.4	Controversy detection.	Yes
MA.5	Automated indicators and key points to assist moderators.	No
MA.6	Event evolution infographic generator.	Yes
RS.1	Different styles/tone of generated reports.	No
RS.2	Abstractive summarization and statistical or analytical data.	Yes
RS.3	Generates reports on the implicit impact of discussions.	No
RS.4	Analytics reports and statistical analysis reports.	No
RS.5	Summary report generator that maintains source links	Yes (partially)
RS.6	Real-time audience feedback analysis.	No
CF.1	Idea filtering and grouping.	Yes

CF.2	Mini-voting mechanism (e.g. majority voting)	No
CF.3	Connection and query external knowledge bases.	Yes
CF.4	Merge and prioritisation of key-points.	No
CF.5	Recommendation of experts according to the topic of deliberation.	No
CF.6	Social media sharing.	No
MPD.1	Support various types of multi-phase deliberations.	Yes
MPD.2	Polling mechanism for advancing deliberation stages.	No
MPD.3	Aggregate survey results.	No
MPD.4	Historical record of deliberation and aggregation of past deliberations.	Yes (partially)
MPD.5	Issue-to-proposal generator and composer of convincing policy change pitch.	No
MPD.6	Prediction of group acceptance and projection into an embedded space of key attributes.	No

The resulting subset spans multiple requirement clusters, with particular emphasis on User Interaction and Engagement (UIE) and Discussion Analysis and Visualization (DAV). Core "strong" requirements retained for the KG context include: a user-friendly interface for exploring network graphs of ideas (UIE.4), clustering discussion data into main viewpoints/arguments (DAV.1), identification of key influential actors (DAV.3), comparison of argument graphs (DAV.4), and an interactive interface for exploring viewpoints (DAV.6). Additional requirements retained as relevant include controversy detection (MA.4), idea filtering and grouping (CF.1), and connections to external knowledge bases (CF.3), among others.

This filtered set functioned as the methodological bridge between ORBIS-level objectives and the KG development work: it defined which system capabilities and interaction goals were considered "in-scope" for the KG platform, before moving to implementation and design decisions in later chapters.

3.7.3.2 Datasets from API call and their structure

From a technical standpoint, ORBIS specifies that the KG operates on structured deliberation data expressed as nodes, relationships, and corresponding properties, typically provided in CSV/JSON format (or via API streams). Expected inputs include deliberation items such as issues,

positions, arguments in favour/against, alongside relevant metadata (e.g., user/profile or ranking-related data).

To support the KG-related work, ORBIS also relies on a broader data infrastructure that includes monitoring and dashboards through Grafana, used to visualise real-time observability data for system components.

In parallel, ORBIS adopts an integration approach where data are crawled from third-party deliberation platforms (including BCause) and made accessible through RESTful API endpoints for distributing analysis outputs.

Within this context, the materials provided for developing the KG platform included:

- Access to a Grafana dashboard where an initial stage of the KG could be inspected, and where datasets for specific discussions could be exported (CSV/JSON).
- Per-discussion datasets structured as two files: Nodes and Edges, reflecting ORBIS's expectation that the KG can be built from entities and relations with associated properties.
- Access to the ORBIS API, which supports retrieval of topic/discussion results and associated outputs through a standardised interface.
- A JSON author/participant metadata file retrieved via the API, containing anonymised identifiers and pseudonyms, with associated timestamps (e.g., creation and last update), used to reference participants without exposing personal identities.

Together, these sources provided both (i) the graph-structured representations needed to model a discussion as a network (nodes/edges) and (ii) the minimal participant metadata necessary to contextualise contributions in an anonymised way, while keeping the methodology aligned with ORBIS's data exchange model (structured exports and API-based access).

3.8 Design Process and Iterative Development

This section describes the design and development process adopted to contribute to ORBIS, with a specific focus on the Knowledge Graph (KG) workstream as a real-world, consortium-driven design setting. The process combined (i) early contextual inquiry on BCause and its reporting layer, (ii) requirement-driven concept development, and (iii) iterative prototyping and implementation under continuous feedback cycles from both users and the ORBIS consortium.

3.8.1 Overall process

The overall process followed a staged-yet-iterative trajectory, where each stage produced intermediate outputs that were repeatedly refined based on feedback and feasibility constraints.

Onboarding and project framing (April 2025).
The collaboration began with an introduction to ORBIS objectives and the role of BCause within the integrated solution. This phase aligned the work with the project's goals (supporting large-scale deliberation with transparent, human-centred AI) and clarified the contribution scope within the visualisation and analytics stream.

Contextual observation of BCause and reporting (Spring–Summer 2025).
The first focus was to understand how BCause structures deliberative discussions and how the report generation mechanism communicates outcomes to end users. This included a design-oriented reading of report components, what information is foregrounded, and where users might face comprehension barriers when AI outputs are shown.
In June I had the opportunity to participate in an ORBIS workshop in Milan with the ORBIS team and some of the ORBIS pilots; I presented some of the concepts that I worked on for the BCause report and aligned with the feedback.

Narrative layer and framing of AI-enhanced components (September 2025).
A second step concentrated on how report content could be made more intelligible through microcopy and framing: clearer component titles, concise descriptions, and explicit indicators that signal when and how AI has been involved. The goal of this step was methodological (improving interpretability and avoiding "expert-only" explanations), without defining or implementing final solutions yet.

Bridging raw outputs to understandable explanations (October 2025).
A key challenge was translating raw outputs from ORBIS pipeline modules (especially Argument Mining) into representations that can be interpreted by non-expert users. This stage informed the logic of explanation and disclosure that later became relevant when designing KG-related communication and interaction patterns.

KG-focused prototyping and development (October 2025 – February 2026).
After being introduced to the initial KG data structure and its early representation (visible through Grafana), the work shifted to designing how graph information could be encoded through visual variables and interactions suitable for a network representation. Development proceeded through multiple iterations, using AI-assisted coding tools (e.g., Gemini and ChatGPT models) to speed up implementation while keeping design intent under human control.

Across all phases, two principles structured the workflow:

- Requirement traceability (ORBIS requirements filtered for KG relevance, combined with literature-derived requirements as a design "blueprint")
- Feedback-in-the-loop iteration, where each version was evaluated against usability, comprehension, and feasibility constraints.

3.8.2 Design Phases

While the process was iterative rather than linear, the work can be described through three recurring design phases that cycled across versions:

- Interpretation and planning: Consolidating inputs from ORBIS documentation, consortium feedback, and user observations into actionable objectives for the next iteration (what should improve, what is out of scope, what is technically feasible given available data and infrastructure).

- Design and implementation: Translating objectives into interface changes (information architecture, visual encodings, microcopy/framing, interactions) and implementing them through prototypes and working builds.

- Evaluation and revision: Testing each iteration with (a) user feedback focused on cognitive load and interpretability, and (b) consortium verification focused on requirement alignment and project integration constraints—then revising the backlog accordingly.

These phases were supported by two continuous activities, described below: sustained user engagement and bi-weekly consortium verification.

3.8.2.1 Sustained user engagement for continuous assessment

User involvement was integrated throughout development, and at the end in a final version Semi-structured interview. The purpose was to ensure that (i) the KG interface remained understandable for non-expert audiences and (ii) the design did not exceed reasonable data literacy demands or cognitive load.

User engagement primarily targeted:

- Comprehension checks (whether users could correctly interpret what the interface shows, especially where AI-enhanced transformations are present),

- Cognitive load monitoring (whether the amount of information, terminology, and interaction complexity remained manageable),

- Barrier detection (moments where users needed "expert knowledge" to proceed, indicating a need for clearer framing or alternative explanations),

- Trust and contestability signals (whether users understood what could be inspected, questioned, or traced back to the discussion data).

Findings from these checks were treated as design constraints: when feedback suggested that the interface demanded too much domain expertise, revisions prioritised clearer wording, stronger framing, and more progressive disclosure of complexity.

3.8.2.2 Bi-weekly verification of how requirement are turned into design KG components with ORBIS consortium

From the beginning of the KG development period, progress was reviewed on a bi-weekly cadence with ORBIS consortium members. These reviews served two methodological functions:

- Requirement-to-design translation control: Each iteration was discussed in terms of which ORBIS-relevant requirements it addressed, what interpretability or accountability goal it supported, and what trade-offs were introduced (e.g., performance constraints, data availability, integration limitations).

- Feasibility and integration alignment: Consortium feedback ensured that design decisions remained consistent with ORBIS technical infrastructure and the BCause ecosystem (available datasets, API access patterns, and reporting workflows), preventing the design framework from drifting into "ideal" solutions that would not be deployable.

This bi-weekly loop functioned as an ongoing governance mechanism: it kept the KG work aligned with ORBIS priorities, while allowing rapid iteration and correction when assumptions (technical or user-facing) did not hold in practice.

4.0
THEORETICAL BACKGROUND. SYSTEMATIC LITERATURE REVIEW FINDINGS

4.1 Literature Landscape

The final review set contains 28 papers and reflects a deliberately interdisciplinary evidence base spanning explainable AI / HCXAI, narrative visualization and data storytelling, and AI-mediated civic participation. In publication terms, the corpus is weighted toward journal venues (18 journal articles), complemented by conference papers (8), and a small number of synthesis-oriented contributions (e.g., surveys and a technical report). Temporally, the landscape is strongly shaped by a recent acceleration: while the corpus includes a small number of foundational anchors (e.g., early work formalizing narrative visualization and visualization rhetoric), more than half of the papers were published from 2024 onwards, indicating both the rapid diffusion of AI-mediated civic tooling and the parallel consolidation of HCXAI as a distinct research agenda.

Across this landscape, three partially decoupled disciplinary streams are visible. First, XAI and HCXAI research contributes conceptual taxonomies, methodological distinctions, and an expanding body of empirical work on how different audiences interpret and use explanations. Large-scale syntheses frame “explainability” as a heterogeneous design space of goals (e.g., accountability, decision support, learning) and techniques (e.g., intrinsic vs post-hoc, local vs global), while also stressing that responsible deployment requires attention to stakeholder needs and context rather than technique alone (Barredo Arrieta et al., 2020). Within this stream, HCXAI is positioned explicitly as a response to technique-centered XAI, shifting emphasis toward sociotechnical settings, personalization, and non-expert interpretation (Ridley, 2025).

Second, narrative visualization and data storytelling research contributes models of how complex information is rendered legible and persuasive through sequencing, annotation, interaction, and genre conventions. A core claim in this stream is that narrative visualizations are hybrid artifacts: they merge communicative intent (“telling a story”) with exploratory affordances typical of information visualization, creating a continuum between author-driven and reader-driven experiences (Segel & Heer, 2010). Complementing this design-space view, visualization rhetoric formalizes how “editorial” choices in narrative graphics can prime particular interpretations: framing is not an accidental byproduct but a predictable effect of rhetorical techniques operating across data selection, visual encoding, textual scaffolding, and interaction design (Jessica Hullman & Nick Diakopoulos, 2011).

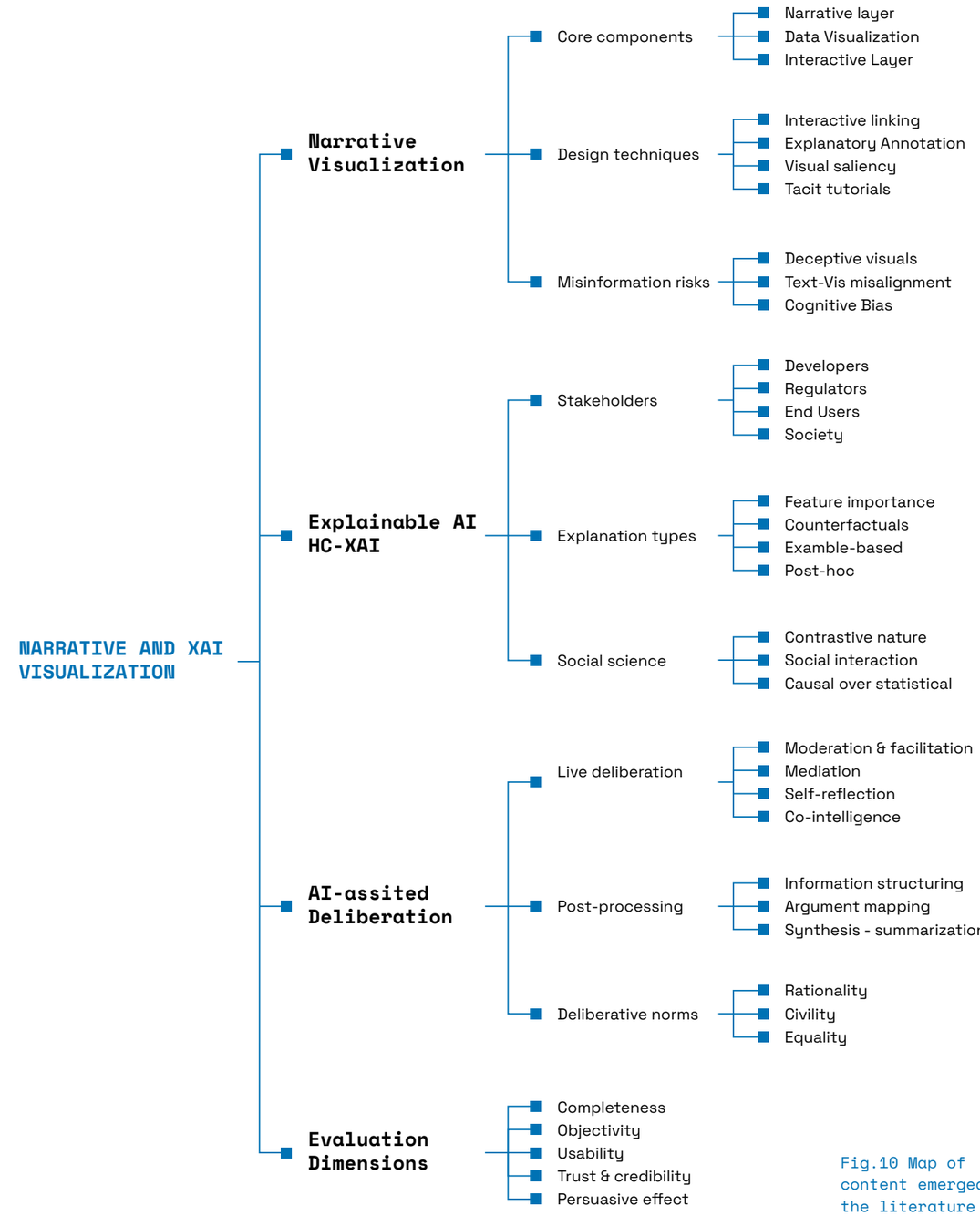


Fig.10 Map of content emerged from the literature

Third, civic-tech and deliberation-oriented research frames AI as an infrastructural mediator of online discourse, often motivated by scale problems (large volumes of citizen input, moderation constraints, low-quality or exclusionary participation). Within this stream, review work maps AI features proposed to support "mass deliberations," foregrounding recurrent challenges such as exclusion, low participation, and difficulties in moderating large-scale discussion (Zeginis et al., 2026). At the same time, normative critique highlights that many AI interventions operationalize a narrow deliberative ideal—especially rationality and civility—as if it were politically neutral; this risks privileging dominant communicative styles and marginalizing forms of participation that do not match those norms (Carstens & Friess, 2024).

A key finding of the landscape mapping is that integration across the three pillars remains limited. Many contributions are strong within one pillar while only weakly engaging the others, and fully "triangular" work is comparatively rare. This fragmentation is not only topical but also methodological: XAI/HCXAI work often evaluates explanations in controlled settings using task-performance measures and self-report constructs; narrative visualization research frequently develops descriptive frameworks and design typologies grounded in case analysis; deliberation research introduces normative criteria (legitimacy, inclusion, equality of voice) that are not straightforwardly reducible to standard usability or comprehension metrics. The result is an evidence base that is rich but difficult to align without an explicit integrative lens.

Within the HCXAI stream specifically, recent synthesis work makes the evaluation problem visible as a central fault line. A survey of user studies for model explanations shows that empirical evaluations span diverse goals and outcome variables (e.g., perceived understanding, trust, task performance), but effects are mixed and highly task-contingent, complicating any "best practice" narrative (Rong et al., 2024). Moreover, empirical evidence suggests that attitudinal trust and behavioral reliance can diverge: explanations may change how much people defer to AI outputs without necessarily changing reported trust, implying that "increasing trust" is an unstable or even misleading evaluation target (Scharowski et al., 2023). This distinction is especially consequential for civic settings, where the design goal is not compliance but calibrated reliance under uncertainty and contestable conditions.

Bridging work that does connect pillars tends to appear in applied, domain-specific systems where explanation is operationalized as interactive sensemaking rather than as a static rationale. For instance, IF-City explicitly

targets intelligibility in fairness-aware urban planning, arguing that fairness tooling should be inclusive of domain experts (not only data scientists) and that domain-specific visualizations are necessary to support iterative, expertise-informed reasoning about inequality (Lyu et al., 2024). Such contributions are important because they treat explanation as part of a workflow of negotiation and revision, closer to deliberative requirements, yet they also illustrate a broader pattern: cross-pillar integration frequently emerges through concrete design problems (planning, policy analysis, platform governance) rather than through shared theory.

4.2 Disciplinary Contributions at the Intersection of XAI, Visualization, and Civic Participation

This subsection situates the SLR corpus across three research lineages—(i) XAI/HCXAI, (ii) information visualization and data storytelling, and (iii) civic participation and democratic deliberation, highlighting what each discipline contributes to the problem of making AI-mediated collective representations legible, contestable, and actionable for non-expert publics. Across the corpus, the convergence is not at the level of "methods," but at the level of human objectives: explanation is treated as a situated communicative act; visualization is treated as a rhetorical and interactional medium for sensemaking; and civic participation research foregrounds legitimacy, inclusion, and the socio-technical conditions under which mediated deliberation can be considered fair.

4.2.1 From XAI to HCXAI: explanation as a situated, plural objective

A first contribution comes from work that reframes "explainability" away from a purely technical property and toward a human-centered, context-dependent practice. A key move in this direction is to treat explanation as grounded in how humans ask and answer "why" questions, where explanations are selective, audience-relative, and often contrastive rather than exhaustive (Tim Miller, 2019). This social-science orientation matters for civic deliberation because public audiences rarely need (or can process) full model internals; instead, they require explanations that help them assess what the system did, why it did it, what would change the outcome, and where the limits and uncertainties are (Tim Miller, 2019).

A second contribution is conceptual: several works argue that "interpretability" is not a single construct and that the literature often

conflates distinct goals (Lipton, 2018). Lipton's critique is particularly relevant for deliberative settings because it challenges the assumption that simply providing an explanation will yield justified trust or appropriate reliance; instead, interpretability claims must be tied to explicit objectives (e.g., contesting decisions, detecting spurious correlations, auditing bias) and to the capacities of particular audiences (Lipton, 2018). Similarly, surveys and taxonomies of XAI consolidate the field around families of approaches and stakeholder needs while emphasizing that responsible deployment requires aligning explanation forms with real-world accountability demands (Barredo Arrieta et al., 2020).

Within this landscape, HCXAI is articulated as a socio-technical response to algorithm-centric XAI, prioritizing user needs, contexts of use, and human evaluation practices (Ridley, 2025). Herrera extends this critique by emphasizing definitional fragmentation in XAI and proposing axes that shift attention from "explaining models" to supporting human-AI collaboration under practical constraints (Herrera, 2025). Together, these accounts position explainability as an interface and interaction problem: the unit of analysis is not only the model, but the explanatory relationship between system, representation, and user.

A third contribution is methodological and empirical: the corpus includes work that systematizes how explanations are evaluated with people and what outcomes are typically measured. A large-scale survey of user studies in XAI synthesizes patterns in study designs, tasks, and evaluation approaches, underscoring that explanation quality cannot be inferred from algorithmic properties alone (Rong et al., 2024). Complementary empirical work operationalizes "comprehensibility" by comparing how participants with different expertise profiles interpret commonly used XAI methods, explicitly treating explainability as a multidisciplinary measurement problem rather than a model-side feature (Bobek et al., 2025). Related experimental evidence examines how explanation styles shape trust and reliance, reinforcing that human-centered explanations can change user behavior and that evaluation must attend to both perceived understanding and downstream reliance decisions (Scharowski et al., 2023).

Finally, civic participation introduces a distinctive evaluative lens: rather than asking only whether explanations increase acceptance, citizen-focused deliberative formats can be used to elicit public preferences about trading accuracy against explainability in consequential decision-making (van der Veer et al., 2021). The HCXAI contribution to the thesis problem is twofold: it supplies (i) conceptual guardrails against monolithic "trust" narratives (Lipton, 2018; Ridley, 2025) and (ii) an empirical agenda for

measuring comprehension and reliance effects under realistic constraints (Bobek et al., 2025; Rong et al., 2024; Scharowski et al., 2023).

4.2.2 Information visualization and data storytelling: rhetoric, interaction, and the construction of meaning

Visualization research contributes a mature language for describing how representations support sensemaking, particularly through task models, interaction patterns, and narrative techniques. Classic visualization taxonomies frame visualization as a tool for cognition oriented toward tasks (e.g., overview and progressive disclosure), which is foundational for designing explainability interfaces that must balance orientation, inspection, and detail under limited attention (Shneiderman, 1996). Building on this, narrative visualization research addresses the tension between author-driven communication and reader-driven exploration, offering a design space of narrative structures and tactics (e.g., annotation, sequencing, highlighting) for turning data into an interpretable account without eliminating the possibility of inquiry (Segel & Heer, 2010).

A central contribution for civic deliberation is the recognition that narrative visualizations are not neutral conduits: they embed rhetorical choices that can systematically shape interpretation. Hullman and Diakopoulos formalize this as "visualization rhetoric," identifying how framing can occur at multiple editorial layers—data selection, visual encoding, annotation, and interactivity—and arguing that these choices can make certain interpretations more probable (Jessica Hullman & Nick Diakopoulos, 2011). This framing perspective directly maps onto deliberative requirements: if an AI-mediated system summarizes, clusters, or aggregates citizen contributions, its communicative choices become a site of power that must be made inspectable and contestable.

The corpus also contributes empirical evidence on the performance effects of storytelling. Experimental work testing data storytelling against more conventional retrieval/analysis setups provides evidence that narrative structure can influence efficiency and effectiveness in information retrieval and insight comprehension (Shao et al., 2024). Design-oriented contributions further support operationalization: a component-based decomposition of narrative visualizations provides a vocabulary for designers to structure data-driven stories as assemblages of reusable parts, enabling more systematic authoring and critique (Borges et al., 2025). Interaction-focused studies in journalistic settings likewise identify how interactive techniques can be layered onto existing visualizations

to promote insight while maintaining engagement, an important bridge between explanatory scaffolding and user agency (Alexandre, 2016).

However, visualization research also foregrounds risks that become acute in democratic contexts. Narrative visualizations can have persuasive effects, shaping attitudes rather than just supporting understanding, which implies that civic explainability must address the ethics of persuasion and the boundary between guidance and manipulation (Chambers & Denham, 2025; Jessica Hullman & Nick Diakopoulos, 2011). Relatedly, work on misinformation in data stories shows that tighter integration of text and visualization can change how audiences evaluate and interpret claims, positioning narrative design as a potential intervention point for robustness against misleading readings (Zheng & Ma, 2022). Finally, an emerging computational thread surveys how foundation models may support the authoring of narrative visualizations—lowering barriers and enabling new workflows—while implicitly raising governance questions about provenance, control, and the accountability of generated narratives (He et al., 2025).

Overall, visualization contributes three assets to the thesis problem: (i) interactional and task-oriented design principles for progressive legibility (Segel & Heer, 2010; Shneiderman, 1996), (ii) a rhetorical theory of how representations guide interpretation (Jessica Hullman & Nick Diakopoulos, 2011), and (iii) empirical grounding that storytelling can measurably affect comprehension and persuasion, demanding explicit ethical and evaluative criteria (Chambers & Denham, 2025; Shao et al., 2024; Zheng & Ma, 2022).

4.2.3 Civic participation and democratic deliberation: legitimacy, inclusion, and infrastructural constraints

A third disciplinary contribution comes from civic participation research that treats communication formats and platform features as determinants of inclusion, deliberative quality, and legitimacy. In open government data dissemination, a study demonstrates that non-expert citizens may prefer data storytelling over portals or dashboard-style communication, and it proposes evaluation emphases—such as completeness and usability—when the goal is actual public understanding rather than expert reuse (Kempeneer et al., 2025). This is a complement to XAI and visualization work: it operationalizes “transparency” as a civic outcome (understanding and potential accountability) rather than an information-release principle.

At the level of participation platforms, several contributions address the problem of scale and information overload. Lightweight visual analytics

tools combine filtering, clustering, and interactive graphs to support exploration of citizen participation content, framing visualization as infrastructural support for navigating large deliberative corpora (Bachiller et al., 2021a). Learning-analytics dashboards similarly aim to support sensemaking in group discussion, treating the “health” of debate and participation patterns as designable, measurable, and improvable (Ullmann et al., 2019). A broader review of AI features for mass deliberation consolidates how AI is being positioned to moderate, summarize, and structure contributions while noting recurring challenges such as participation inequities and moderation burdens (Zeginis et al., 2026).

A parallel HCI thread studies micro-level interventions on deliberation quality. Interface nudges that scaffold self-reflection are empirically tested as a way to enhance deliberativeness on online platforms, foregrounding the role of interaction design in shaping discourse norms (Yeo et al., 2024). Participatory design research further argues that deliberating “with AI” should itself be a stakeholder process: AI-supported decision-making systems must be designed through structured deliberation that surfaces values and contestation around objectives, constraints, and impacts (Zhang et al., 2023). At the same time, normative critiques warn that AI tools for online discourse often operationalize deliberation through narrow norms of rationality and civility, potentially privileging already-dominant styles of participation and marginalizing other legitimate forms of expression (Carstens & Friess, 2024). This line of work directly informs deliberative requirements such as plurality preservation and the avoidance of epistemic closure.

Finally, the corpus includes systems-level demonstrations of AI mediation in democratic deliberation. Large-scale experiments on AI-mediated group statement generation show that an LLM-based “mediator” can produce statements that participants endorse and that are judged as high quality along dimensions such as clarity and perceived fairness, while also incorporating minority critiques during iterative revision (Tessler et al., 2024). This is directly relevant for the thesis problem because it instantiates an AI-mediated deliberative pipeline whose outputs are collective representations—and thus raises the question of what “evidence traces,” aggregation logic, and exclusions must be made legible for legitimacy and contestability.

Two additional contributions concretely bridge civic participation with XAI and visualization. In participatory fairness and planning, an “intelligible” visual analytics tool for urban planning integrates causal, contrastive, and counterfactual explanation modes to help domain experts perceive

inequality, attribute sources, and explore mitigation actions—positioning explanation as an interaction loop rather than a static justification (Lyu et al., 2024). In policy deliberation platforms, LLMs are used to structure discourse into positions and argument maps while explicitly linking AI-generated structure to human-authored statements, thereby proposing a traceable bridge between computational organization and citizen voice (Bhatia & Sukthankar, 2025). These works illustrate a key design direction for civic HCXAI: coupling interpretive structure with inspectable provenance.

4.2.4 Synthesis: what the intersection produces

The corpus shows that progress at the intersection depends on integrating three kinds of disciplinary knowledge. HCXAI contributes a theory of explanation as social, goal-relative, and evaluable with users—while cautioning against collapsing explainability into "trust" (Lipton, 2018; Ridley, 2025; Rong et al., 2024; Tim Miller, 2019). Visualization contributes a theory of sensemaking through progressive disclosure and interaction, alongside an explicit account of rhetoric and framing that is indispensable when collective representations can shape political interpretation (Jessica Hullman & Nick Diakopoulos, 2011; Segel & Heer, 2010; Shneiderman, 1996). Civic participation research contributes legitimacy-driven evaluation criteria, evidence that storytelling and interface scaffolds shape public understanding, and warnings that AI mediation can reproduce power through the norms it encodes (Carstens & Friess, 2024; Kempeneer et al., 2025; Tessler et al., 2024; Yeo et al., 2024).

This alignment sets up the next section by making visible a shared problem statement across fields: AI-mediated deliberation requires representational forms that help publics understand and critique how collective claims are produced, without collapsing plurality into a single authoritative narrative.

4.3 Thematic Synthesis and Meta-Themes

This section synthesizes the SLR corpus through four cross-cutting meta-themes that translate heterogeneous contributions (XAI/HCXAI, narrative visualization/storytelling, and democratic deliberation) into design-relevant knowledge. Rather than summarizing papers sequentially, each theme consolidates recurring mechanisms, risks, and evaluation implications, while acknowledging that individual papers frequently contribute to multiple themes. The aim is to surface what must be made communicatively legible in AI-mediated deliberation systems, how sensemaking can be scaffolded

without collapsing disagreement, which interactional affordances enable contestability, and how to evaluate civic explainability beyond trust-centric metrics.

4.3.1 Transformation transparency

In the corpus, "transformation transparency" concerns the legibility of how AI-mediated deliberative pipelines turn dispersed citizen inputs into collective representations (e.g., clusters, positions, summaries, "group statements"), including the normative choices embedded in moderation and structuring. This is distinct from model-centric explainability: the object of explanation is the mediation process (selection, aggregation, re-expression, and representation), not only the model's internal logic.

Normative assumptions as part of the pipeline

A foundational move in this theme is to treat the operationalization of "deliberative quality" as a transparency problem. Carstens and Friess argue that AI tools for online discourse often encode deliberative norms—especially rationality and civility—through detectable proxies (e.g., argument mining, markers of incivility), thereby simplifying complex deliberative dimensions and shifting attention toward an input–output framing of "bias" that can obscure deeper normative interference (Carstens & Friess, 2024). From a transformation-transparency perspective, this implies that civic explainability must make visible not only what the system outputs, but also what it treats as a problem (e.g., incivility), which indicators it uses, and what forms of participation may be disadvantaged by those operational choices (Carstens & Friess, 2024).

Traceability from collective representations back to human contributions

Several intersection papers position traceability as an explicit design choice: outputs become more legitimate when users can see the human evidence they are built from. Bhatia and Sukthankar propose using LLMs to generate higher-level "positions" and structure Polis forum discourse into argument maps, while explicitly grounding each AI-generated position in supporting human statements (Bhatia & Sukthankar, 2025). This "support-by-statements" strategy functions as a concrete provenance link: the representation is interpretable not only as an abstract summary but as a structured index into citizen voice (Bhatia & Sukthankar, 2025).

A parallel logic appears in AI-mediated deliberation experiments where the system iteratively generates group statements and revises them in response to participant feedback; the reported design includes

incorporating minority critiques into revised statements, emphasizing that mediation is a process with revision history rather than a one-shot synthesis (Tessler et al., 2024). Even when the primary contribution is efficacy (e.g., quality and perceived fairness of mediated statements), this work foregrounds an important transparency requirement for civic settings: users should be able to inspect how a collective statement emerged over iterations, including what feedback was integrated and what was not (Tessler et al., 2024).

Making aggregation and feature-level AI support explicit

A review of AI features for mass deliberation consolidates recurring pipeline functions—moderation, summarization, processing deliberation data, and participation enhancement—while also noting that integration raises risks around fairness and transparency (Zeginis et al., 2026). In the context of transformation transparency, this reinforces that "AI for deliberation" is typically pipeline-shaped: multiple stages transform discourse, and transparency must address the cumulative effects of these stages (Zeginis et al., 2026). A design implication is that publics may need legibility over (at minimum) the sequence of stages, the criteria used at each stage (e.g., what counts as low-quality or off-topic), and the scope conditions under which the system's representation is meant to hold.

Transparency for algorithmic reasoning about public consequences

Transformation transparency also includes making consequential analytic transformations legible to non-ML stakeholders. Lyu et al. argue that fairness-oriented visual analytics tools have often targeted data scientists, whereas fairness must involve domain experts; they propose a framework and tool that support domain reasoning through causal attribution ("Why"), contrastive ("Why not"), and counterfactual ("What if/how to") explanation modes in allocation/planning contexts (Lyu et al., 2024). Here, transparency is operationalized as workflow intelligibility: users can perceive inequality patterns, attribute sources, and explore mitigation through interactive simulation and recommendations, rather than receiving opaque optimization outputs (Lyu et al., 2022).

Interface strategies for making transformation inspectable under cognitive constraints

Multiple civic-analytics contributions instantiate transparency as navigable interfaces over complex discourse corpora. Bachiller et al. present a lightweight web application for citizen participation content that combines filters (e.g., date, location, category) with interactive graphs to support exploration and analysis in an "easy and clear way," explicitly motivated by

information overload in e-participation platforms (Bachiller et al., 2021a). Ullmann et al. similarly develop a visualization dashboard for "contested collective intelligence" to support engagement and sensemaking of complex group discussions, positioning learning analytics visualizations as aids for users who must stay abreast of collectively generated information (Ullmann et al., 2019). In both cases, transparency is enacted as selectable views and interactive reduction of complexity—a reminder that "more information" is not the same as "more legible" (Bachiller et al., 2021b; Ullmann et al., 2019).

Transparency challenges introduced by AI-assisted narrative generation

A further, increasingly relevant, pipeline layer concerns the use of foundation models to craft narrative visualizations. He et al. survey how foundation models can support stages of narrative visualization authoring and point to the importance of "displaying the basis and context of AI-generated content" to make it more trustworthy (He et al., 2025). Even when the setting is not deliberation per se, the implication transfers directly: if AI helps generate civic-facing narratives from deliberative data, transformation transparency must include provenance and contextual grounding for generated claims and story structure (He et al., 2025).

Tab. 04 Papers mapped to MT1 Transformation Transparency

Paper ID	Authors	Evidence contribution
P05	(He et al., 2025)	contributes a survey-level pipeline view of AI-assisted narrative visualization authoring and explicitly highlights the need to show the basis/context of AI-generated content for trustworthiness.
P09	(Carstens & Friess, 2024)	contributes a normative critique showing that AI discourse tools often operationalize rationality/civility through simplified proxies, motivating transparency of embedded deliberative norms and their distributive effects.
P12	(Zeginis et al., 2026)	contributes a feature-level map of AI functions in mass deliberation (moderation/summarization/processing) and flags fairness/transparency risks, reinforcing that transparency must cover multi-stage pipelines.
P14	(Bhatia and Sukthankar, 2025)	contributes a traceability pattern by linking LLM-generated positions/argument maps to supporting human statements in Polis discourse.
P15	(Bachiller et al., 2021)	contributes an interface pattern for inspectability under overload via lightweight filters and interactive graphs over participation content.

P16	(Ullmann et al., 2019)	contributes a dashboard approach for sensemaking in contested group discussion, framing transparency as navigable learning-analytics views.
P19	(Lyu et al., 2022)	contributes a workflow intelligibility model for fairness-aware planning using "Why/Why-not/What-if/How-to" explanation modes paired with interactive simulation and recommendations.
P21	(Tessler et al., 2024)	contributes an AI-mediation pipeline exemplar (iterative group-statement generation) and reports integration of minority critiques into revised statements, motivating transparency over revision logic and evidential incorporation.

4.3.2 Plurality-preserving sensemaking

Plurality-preserving sensemaking concerns how narrative/visual representations can make complex civic material intelligible while maintaining the legitimacy of disagreement. In the reviewed corpus, "sensemaking" is not treated as a purely cognitive task (helping users understand), but as an epistemic and political requirement: collective representations should support comprehension and orientation without prematurely converging on a single interpretation or suppressing minority positions. This requirement becomes more acute in AI-mediated deliberation, where pipelines restructure discourse into intermediate outputs (e.g., clusters, "insights," summaries, argument maps) that shape what becomes visible and actionable.

Plurality preservation as a design problem of balancing guidance and agency

Narrative visualization has been described as occupying a spectrum between author-driven sequencing and reader-driven exploration, where the core design challenge is to provide orientation without removing the audience's capacity to interrogate or reframe the data (Segel & Heer, 2010). Building on this, a component-level characterization of narrative genres makes explicit how narrative structure is assembled from reusable parts (e.g., visual representation, interaction techniques, degrees of freedom of exploration, and narrative approach), implying that plurality can be supported by selecting components that maintain exploration "escape hatches" rather than enforcing a single path (Borges et al., 2025). Empirical analysis of interaction in existing visualizations further supports the view that interaction is not merely decorative but can be used to promote insight by enabling users to ask their own questions through mechanisms

such as filtering, selection, and details-on-demand (Alexandre, 2016). These contributions position plurality-preserving sensemaking as a matter of narrative architecture: scaffolding should guide attention and reduce noise while retaining the capacity for alternative readings and self-directed inquiry.

Cognitive scaffolds under heterogeneous literacy, highlighting trade-offs between efficiency and epistemic accessibility

Data storytelling elements (e.g., explanatory titles, annotations, visual emphasis, and narrative cues) can improve effectiveness in comprehension-oriented tasks, including accuracy, even when they may not optimize speed in point-retrieval tasks (Shao et al., 2024). These benefits do not straightforwardly concentrate among low-literacy participants, complicating a common assumption that storytelling is inherently compensatory for limited visualization literacy (Shao et al., 2024). Complementary evidence comes from work on text–visualization integration in data stories, which investigates whether enhanced coupling between narrative text and visual evidence can reduce misinterpretations and mitigate misinformation effects—suggesting that how text points to evidence, and how evidence is staged over time, materially shapes interpretive outcomes (Zheng & Ma, 2022). In public-sector communication, a field-lab redesign of open government data dissemination shows that non-expert citizens often prefer narrative "scrolly" formats over portals and dashboards, because storytelling provides context and a "bigger picture" needed to interpret complex policy problems (Kempeneer et al., 2025). Across these studies, the plurality-preserving implication is that scaffolding should be conceived as progressive support (helping diverse audiences reach interpretive competence) rather than interpretive substitution (replacing audience judgement with a pre-interpreted storyline).

Narrative visualization is rhetorically consequential, and therefore must be designed for contestability rather than treated as neutral transmission

Visualization rhetoric research argues that narrative visualizations can frame interpretation through selective emphasis, sequencing, and contextual cues; such framing effects are not incidental, but structurally embedded in narrative design (Jessica Hullman & Nick Diakopoulos, 2011). This resonates with evidence from open government data storytelling, where "objectivity" is problematized: a neutral presentation is treated as unattainable in practice, and the proposed alternative is to keep governmental narratives open to discussion and challenge, reconfiguring objectivity as a co-creative process between institutions and citizens

(Kempeneer et al., 2025). For civic deliberation, these contributions shift the design target from “a persuasive or coherent story” to “a navigable and accountable representation,” where rhetorical choices are expected to be questioned and where alternative interpretations can be articulated against the same evidentiary base.

Plurality preservation in concrete representational forms for disagreement, trade-offs, and counterfactual exploration

In AI-mediated structuring of deliberation, LLM-based pipelines can generate higher-level “positions” and organize discourse into argument maps while explicitly grounding AI-generated positions in human-authored statements and leveraging platform voting to estimate acceptance (Bhatia & Sukthankar, 2025). This approach foregrounds plurality as a mapping problem—how to summarize without severing links to diverse contributions—while also illustrating an unresolved tension: prioritizing high-consensus “insights” can risk treating refutation and minority viewpoints as secondary, even when those are deliberatively valuable (Bhatia & Sukthankar, 2025). In a different domain, intelligible fairness-aware city planning frames plurality as the ability to explore distributive impacts across stakeholders and constraints through explanations that include causal attribution, contrastive reasoning, and counterfactual “what-if/how-to” exploration, enabling domain experts to understand and revise allocation decisions iteratively (Lyu et al., 2024). Together, these works suggest that plurality-preserving sensemaking benefits from representations that keep disagreement and alternatives structurally available—through explicit support links, stakeholder-differentiated views, and counterfactual navigation—rather than compressing the space of possible interpretations into a single synthesized outcome.

How plurality is affected when narrative construction becomes partially automated

A survey of foundation models for narrative visualization argues that generative systems can support tasks across the narrative visualization workflow (from data transformation to story drafting and refinement), but also raises the stakes of controllability, reliability, and human oversight when models participate in authorial decisions (He et al., 2025). In the context of deliberation summaries and “positions,” this reinforces a core requirement of plurality-preserving sensemaking: AI assistance should not only produce fluent narratives, but must do so under constraints that maintain traceability to human inputs and preserve users’ ability to contest and reframe the representation (Bhatia & Sukthankar, 2025; He et al., 2025). Across the meta-theme, the literature converges on a design-oriented conclusion aligned with RQ2: narrative/visual scaffolding is valuable when

it functions as an access mechanism to plural civic meaning, not as a mechanism for epistemic closure.

Tab. 05 Papers mapped to MT2 Plurality-preserving Sensemaking

Paper ID	Authors	Evidence contribution
P02	(Shao et al., 2024)	Quantifies how storytelling elements affect efficiency/effectiveness; shows scaffolds can improve accuracy/comprehension but do not automatically compensate for low visualization literacy.
P03	(Borges et al., 2025)	Provides a component-level “structure” of narrative visualization genres, supporting deliberate design choices that preserve exploration and alternative readings.
P04	(Zheng and Ma, 2022)	Examines how tighter text–visual integration shapes interpretation in data stories under misinformation pressure, highlighting design sensitivity of narrative coupling.
P05	(He et al., 2025)	Maps how foundation models can support narrative visualization workflows while foregrounding requirements for control/oversight—critical when narratives mediate civic meaning.
P07	(Segel and Heer, 2010)	Establishes the author-driven vs reader-driven spectrum and genre patterns, framing plurality as a balance between guidance and audience agency.
P08	(Alexandre, 2016)	Shows how interaction added to existing visualizations can promote insight, supporting plurality through user-driven inquiry and exploration.
P14	(Bhatia and Sukthankar, 2025)	Proposes LLM-based structuring of Polis discourse into argument maps grounded in human statements and voting-derived acceptance, raising issues of how disagreement is represented.
P19	(Lyu et al., 2024)	Operationalizes plural exploration via intelligible fairness planning using causal/contrastive/counterfactual explanation types and iterative revision of trade-offs.
P20	(Kempeneer et al., 2025)	Empirically demonstrates that non-expert citizens prefer narrative formats for policy understanding and argues for multiplicity of narratives and co-creative “objectivity.”
P24	(Hullman and Diakopoulos, 2011)	Theorizes narrative visualization as rhetorical framing, motivating contestable and reflexive narrative designs to avoid epistemic closure.

4.3.3 Interaction, agency and contestability mechanisms

From legibility to governability

Across the corpus, interaction is not framed as an ergonomic layer for accessing information, but as the key means through which publics and stakeholders can exercise agency over AI-mediated collective representations. This shifts the role of explainability from a post-hoc justification ("here is why the system produced this output") to a practical capacity to inspect, question, and intervene in the representational pipeline. Narrative visualization research already describes interactivity as a balancing act between author-imposed flow and reader-driven discovery (Segel & Heer, 2010). In AI-mediated deliberation, this balance becomes politically consequential: the interface effectively allocates interpretive power—who can probe assumptions, surface exceptions, and challenge outcomes.

Interaction budgets and non-expert constraints

Several papers foreground a core usability tension: civic participants are rarely "visual analytics experts," yet are asked to make sense of complex debate dynamics and model-driven aggregations. Ullmann et al. (2019) explicitly problematize this gap and empirically investigate whether relatively inexperienced users can complete debate-related tasks with alternative visual representations of deliberation content. Their discussion also highlights a known limit of graph-based representations: as participation and discourse ontologies scale, argument maps risk becoming too unwieldy to remain usable, motivating the need for alternative visual encodings and carefully designed interaction strategies (Ullmann et al., 2019). This resonates with interaction choices in narrative visualization, where interactivity is often staged (e.g., constrained exploration after narrative segments) to prevent users from getting lost while still enabling verification and deeper inspection (Segel & Heer, 2010). In deliberation settings, such "staged" interaction can be read as a cognitive scaffolding strategy that preserves access while reducing misinterpretation risks.

Pipeline agency through participatory model-building

A stronger form of agency emerges when interaction is designed to let stakeholders engage in the construction of the model itself. Zhang et al. (2023) propose a workflow in which stakeholders create and evaluate ML models to surface organizational patterns, then deliberate about how decision-making should change. Their web tool structures participation along four recognizable ML pipeline stages (data exploration, feature selection, model training, evaluation), and uses the resulting models as

"boundary objects" for deliberation between decision-makers and decision-subjects (Zhang et al., 2023). Crucially, the tool is not positioned as producing a final "correct" model, but as enabling stakeholders to articulate and negotiate normative values embedded in features, metrics, and fairness criteria—i.e., turning interaction into a mechanism for procedural contestation over how collective judgments are operationalized.

Critique loops as a procedural slot for contestation

Contestability can also be operationalized at the level of collective outputs through explicit critique-and-revision loops. Tessler et al. (2024) design an AI mediator ("Habermas Machine") that generates candidate group statements from participants' private opinions and then revises statements based on participants' written critiques across rounds. The interaction design is structurally important: critique is not an optional comment layer but a defined phase in the deliberation protocol, enabling dissent to be registered and translated into revised collective text (Tessler et al., 2024). In this framing, interaction is inseparable from legitimacy: the "group statement" is not presented for acceptance but is iteratively negotiated through structured opportunities to object and refine.

Action-oriented explanation: why, why-not, what-if, how-to

Among the clearest operationalizations of contestability in the corpus is the explicit mapping of explanation types to user questions and possible interventions. Lyu et al. (2024) propose the IF-Alloc framework and instantiate it as IF-City, an interactive visual tool for fairness-aware city planning. The framework explicitly integrates causal attribution ("why"), contrastive explanations ("why not"), counterfactual reasoning ("what if"), and actionable guidance ("how to") to help domain experts perceive inequality, diagnose sources, and explore mitigation strategies through simulation and constraint-satisfying recommendations (Lyu et al., 2024). While the application domain is urban planning, the interaction logic generalizes to deliberative pipelines: contestability is strengthened when explanations are coupled to actions users can take (e.g., explore alternatives, identify constraints, test trade-offs), rather than remaining purely descriptive.

Co-production interactions in civic contexts

Agency is also expressed through lighter-weight interaction mechanisms that help participants navigate large-scale civic datasets and meaningfully contribute. Bachiller et al. (2021) develop an interactive data mining tool for citizen participation content that supports exploratory filtering across time, geography, categories/topics, and clustered "communities of interest." They add a co-production-oriented feature: users can submit

a prospective new proposal and retrieve similar existing proposals to avoid duplication and potentially merge contributions (Bachiller et al., 2021b). Even without an explicit XAI framing, this interaction functions as a pragmatic contestability mechanism at the content level—citizens can check whether the system's representation of the issue space already includes related contributions and can adjust their participation accordingly.

Micro-interactions that shape agency upstream

Finally, interaction can be targeted not only at reading or contesting outputs, but at shaping the quality of inputs before aggregation and summarization. Yeo et al. (2024) design "self-reflection nudges" embedded in the comment-writing interface of online deliberation platforms, using LLM-powered reflective prompts (e.g., persona-based, temporal, storytelling-oriented) to strengthen internal reflection and deliberative quality (Yeo et al., 2024). This positions agency as something supported upstream: interaction scaffolds the participant's own reasoning and perspective-taking, which then propagates into downstream representations.

Tab. 06 Papers mapped to MT3 Interaction, agency and contestability mechanisms

Paper ID	Authors	Evidence contribution
P07	(Segel and Heer, 2010)	Interaction patterns that balance author-driven flow and reader-driven exploration; staged interactivity as a design strategy for comprehension and verification.
P08	(Alexandre, 2016)	Interaction techniques in narrative visualization as drivers of insight and user control during exploration.
P15	(Bachiller et al., 2021)	Filter-driven exploration of civic corpora; clustering + similarity retrieval as citizen-facing co-production/anti-duplication agency.
P16	(Ullmann et al., 2019)	Usability constraints for deliberation analytics; non-expert limitations; risks of "clumsy" graph representations at scale; need for alternative representations and task-oriented interaction.
P10	Zhang et al. (2023)	Participatory interaction over ML pipeline stages; stakeholders deliberate over features/metrics/fairness; models as boundary objects for procedural contestation.
P11	Yeo et al. (2024)	Interface nudges that scaffold reflection during contribution; agency shaped upstream at the point of writing.
P19	(Lyu et al., 2024)	Contestability via interactive explanation modes (why/why-not/what-if/how-to) tied to mitigation actions and constraint-aware recommendations.
P21	(Tessler et al., 2024)	Structured critique loops for revising collective outputs; contestation embedded as a phase in the protocol.

4.3.4 Evaluating democratic adequacy beyond trust

From "does it increase trust?" to "is it democratically adequate?"

Across the corpus, evaluation is repeatedly problematized as the weak link of XAI-in-use, particularly when systems mediate civic meaning. Standard XAI evaluation often defaults to self-report constructs (especially trust) or narrow task-performance proxies, but several contributions argue that these are insufficient for deliberative settings where legitimacy, inclusion, and contestability are core requirements. Foundational critiques emphasize that "interpretability" is not a single measurable property and that evaluations must be aligned to explicit objectives rather than treated as generic improvements (Barredo Arrieta et al., 2020; Lipton, 2018). A social-science grounding of explanation further frames evaluation as inherently relational and audience-dependent: explanations are selected to answer particular "why" questions, and their adequacy depends on what the recipient needs to do with them (Tim Miller, 2019). From this perspective, democratic adequacy requires moving beyond trust as a proxy outcome toward multi-criterion evaluation linked to comprehension, calibrated reliance, legitimacy, and contestability.

Trust is not a stable proxy for appropriate reliance

One of the clearest empirical messages in the theme is that trust and reliance should not be conflated. Work explicitly investigating human-centered explanations shows that increased transparency does not necessarily change self-reported trust, while behavioral reliance can change (and in task-dependent ways), indicating that "trust gains" are an unreliable summary of explanation effects (Scharowski et al., 2023). This aligns with broader discussions in HCXAI that emphasize calibration—whether users rely on the system appropriately given its limits—rather than indiscriminate increases in trust (Ridley, 2025; Rong et al., 2024). In deliberation, this matters because the goal is not compliance with AI-mediated summaries but the capacity to critically assess collective representations while avoiding both overreliance (epistemic closure) and disuse (dismissal of potentially helpful structuring).

Comprehension depends on literacy, domain knowledge, and evaluation design

Several papers treat "understanding" as a measurable, heterogeneous outcome, shaped by data literacy and domain expertise. A comprehensive empirical study of explainability finds that comprehension varies significantly across user groups and is influenced by both data literacy and domain-specific knowledge, implying that evaluation must segment

users and avoid assuming a generic “public” (Bobek et al., 2025). At the synthesis level, surveys of XAI user studies highlight methodological heterogeneity and the difficulty of generalizing explanation effects across tasks, audiences, and explanation types, reinforcing the need for explicitly defined target populations and outcomes (Rong et al., 2024). HCXAI reviews similarly argue that evaluation should be designed around human goals and contexts of use rather than model properties alone (Ridley, 2025), while broader reflections propose that the field should orient evaluation toward human–AI collaboration rather than technique comparisons detached from real settings (Herrera, 2025).

Ecological validity: tasks, stakes, and civic contexts change what “good” means

A recurring cross-paper implication is that evaluation outcomes shift with stakes and context. Citizen-jury evidence shows that preferences between accuracy and explainability are not fixed: participants weighed explainability differently across domains, favoring accuracy more strongly in healthcare scenarios while valuing explainability equally or more in non-healthcare contexts (van der Veer et al., 2021). This strongly supports the RQ3 framing: evaluation must be sensitive to the civic domain, the perceived consequences, and the decision ecology rather than treating “explainability vs accuracy” as an abstract trade-off. In parallel, ethical analysis of AI interference in online discourse highlights that evaluative framing itself can be normatively loaded: systems that operationalize rationality and civility through simplified proxies may privilege particular groups and communicative styles, so evaluation must include distributive and ethical dimensions, not only usability or perceived helpfulness (Carstens & Friess, 2024).

Legitimacy and representativeness as measurable outcomes

Democratic adequacy foregrounds legitimacy-related evaluation targets that are often absent in standard XAI studies. Large-scale AI-mediated deliberation experiments explicitly evaluate whether mediated group statements represent viewpoints fairly, incorporate minority perspectives, and are preferred by participants compared to human mediation; they additionally report ratings from external judges on dimensions such as quality, clarity, informativeness, and perceived fairness (Tessler et al., 2024). This contribution is important methodologically: it demonstrates evaluation designs that measure the quality and fairness of collective outputs rather than only individual-level trust or understanding, and it treats minority representation and bias risks as explicit research questions (Tessler et al., 2024). In civic communication contexts, open government data dissemination work similarly positions evaluation around

citizen understanding and usability, showing that narrative storytelling formats can increase comprehension relative to portals and dashboards and arguing for narrative multiplicity rather than a single authoritative “objective” storyline (Kempeneer et al., 2025). Together, these works suggest that democratic adequacy can be evaluated through criteria such as representational coverage, minority inclusion, perceived fairness, and usability for non-expert sensemaking.

Contestability and agency as evaluative criteria, not only design goals

Several papers imply that democratic adequacy includes the user’s capacity to challenge, revise, or steer system outputs. Participatory AI design work frames evaluation as a deliberative process where stakeholders build and critique models and deliberate about how decision-making should change, shifting the evaluative focus from “explanation acceptance” to procedural legitimacy and value negotiation (Zhang et al., 2023). In fairness-aware planning, interactive explanation modes (including causal, contrastive, and counterfactual “what-if/how-to” reasoning) are used to support mitigation-oriented exploration of trade-offs, positioning evaluation around whether stakeholders can diagnose issues and explore actionable alternatives (Lyu et al., 2024). In online deliberation interfaces, reflection nudges are evaluated as interventions to enhance deliberativeness at the point of contribution, linking evaluation to discourse quality rather than trust in AI (Yeo et al., 2024). At a landscape level, a review of AI features for mass deliberation underlines that fairness, transparency, and governance concerns are central when AI is integrated into deliberative pipelines, implying that evaluation should extend to platform-level impacts (Zeginis et al., 2026).

Rhetorical effects and persuasion must be explicitly measured in civic settings

Because civic representations are intrinsically rhetorical, the corpus includes work that treats persuasion and framing as measurable—and ethically relevant—outcomes. Narrative visualization studies empirically measure persuasive effects of rhetorical techniques, reminding that “effective” stories may shape attitudes rather than only improve understanding (Chambers & Denham, 2025). Complementary work on misinformation in data stories examines how text–visual integration can change interpretation and misinterpretation risks (Zheng & Ma, 2022). These contributions expand democratic adequacy to include epistemic safeguards: evaluation should test whether narrative/visual scaffolds support comprehension without unduly steering beliefs or suppressing alternative interpretations.

The meta-theme converges on a multi-dimensional evaluation stance: democratic adequacy should be assessed through (i) comprehension and mental-model quality for target publics, (ii) calibrated reliance under uncertainty, (iii) legitimacy-related outcomes such as representational fairness and minority inclusion, and (iv) contestability/recourse capacities enabled by interaction—while explicitly accounting for cognitive constraints and rhetorical side effects (Kempeneer et al., 2025; Lipton, 2018; Rong et al., 2024; Scharowski et al., 2023; Tessler et al., 2024; Tim Miller, 2019).

Tab. 07 Papers mapped to MT4 Evaluating democratic adequacy beyond trust

Paper ID	Authors	Evidence contribution
P01	(Bobek et al., 2025)	Empirically shows that explainability comprehension varies across user groups and depends on data literacy and domain knowledge; motivates segmented, audience-specific evaluation designs.
P06	(Chambers and Denham, 2025)	Measures persuasive effects of narrative visualization rhetoric; frames persuasion as an evaluable (and ethically relevant) outcome.
P09	(Carstens and Friess, 2024)	Argues AI discourse interventions encode and privilege specific deliberative norms; implies evaluation must include ethical/distributive impacts, not only usability.
P10	(Zhang et al., 2023)	Positions evaluation as stakeholder deliberation over ML pipeline choices (features, metrics, trade-offs); shifts adequacy toward procedural legitimacy and value negotiation.
P11	(Yeo et al., 2024)	Evaluates reflection nudges as interventions to improve deliberativeness; operationalizes evaluation in terms of discourse quality rather than trust.
P12	(Zeginis et al., 2026)	Reviews AI features for mass deliberation and associated challenges; supports platform-level evaluation attention (fairness/transparency/governance risks).
P13	(van der Veer et al., 2021)	Uses citizens' juries to evaluate accuracy–explainability trade-offs; shows preferences are context- and stakes-dependent, supporting domain-sensitive evaluation.
P17	(Herrera, 2025)	Synthesizes critiques of XAI and argues for a shift toward human–AI collaboration; reframes evaluation targets beyond technique-centric measures.
P18	(Scharowski et al., 2023)	Empirically separates trust and reliance; shows explanation effects can be task-dependent and that reliance can change without corresponding trust changes, motivating calibration-focused evaluation.
P19	(Lyu et al., 2024)	Evaluates intelligible fairness planning via interactive causal/contrastive/counterfactual explanations; positions adequacy around diagnosis and actionability (mitigation exploration).
P20	(Kempeneer et al., 2025)	Field-lab evidence that storytelling improves citizen understanding of open government data; proposes usability/completeness and narrative multiplicity as civic evaluation criteria.
P21	(Tessler et al., 2024)	Evaluates AI-mediated deliberation through participant preference and external-judge ratings (quality/clarity/informativeness/perceived fairness) and minority-representation checks; operationalizes legitimacy-oriented outcomes.
P22	(Ridley, 2025)	Reviews HCXAI and emphasizes human goals and socio-technical contexts; supports evaluation beyond model properties toward situated human outcomes.
P23	(Rong et al., 2024)	Surveys user studies of model explanations; documents heterogeneity of evaluation practices and supports multi-outcome, context-aware measurement strategies.
P25	(Lipton, 2018)	Critiques interpretability as a conflated construct; argues evaluation must be tied to explicit objectives rather than generic "more interpretable" claims.
P27	(Barredo Arrieta et al., 2020)	Consolidates XAI methods and stakeholder motivations; frames evaluation as alignment between explanation type, user needs, and accountability goals.
P28	(Miller, 2019)	Grounds explanation in social-science theory (contrastive, selective, audience-relative); implies evaluation must consider the questions explanations answer and for whom.

4.4 Current Gaps and Design Opportunities

This section articulates the main gaps that emerge from the SLR corpus and translates them into communication-design opportunities for HCXAI in AI-mediated civic deliberation. The aim is not to claim that the literature lacks methods, but that it often leaves under-specified how methods should be rendered legible and contestable for non-expert publics, how plurality can be preserved under summarization and structuring pressures, and how evaluation can be aligned to democratic adequacy rather than trust-centric proxies (Barredo Arrieta et al., 2020; Carstens & Friess, 2024; Lipton, 2018; Rong et al., 2024; Tim Miller, 2019).

4.4.1 Gap A: From model explainability to pipeline accountability

A recurring limitation across the corpus is that explainability is frequently framed at the level of model logic, while the deliberative mediation pipeline—the end-to-end sequence that transforms citizen contributions into collective representations—remains only partially legible. Yet in civic settings, legitimacy and contestability hinge on whether publics can trace how outputs were produced: which inputs were included, how they were aggregated, which assumptions guided the structuring, what was excluded, and where uncertainty and scope limits apply (Carstens & Friess, 2024; Zeginis et al., 2026). Even when systems explicitly link outputs to evidence, the pipeline remains complex: LLM-based discourse structuring pipelines create intermediate artefacts (topics, positions/insights, support relations, acceptance estimates) that involve multiple algorithmic decisions and heuristics that are rarely exposed as a coherent "transformation story" (Bhatia & Sukthankar, 2025). Similarly, AI-mediated deliberation work that produces group statements at scale shows that the mediation process can affect agreement and minority/majority dynamics, underscoring that the pipeline itself is politically consequential and therefore a target of accountability (Tessler et al., 2024). At the same time, the inspection of the pipeline should be "on demand" and not something exposed to everyone regardless of the context and the intention of the user.

Without pipeline-level transparency, civic users may be asked to accept or act on collective representations without the means to interrogate provenance, exclusions, and value-laden operationalizations of "deliberative quality". Normative critiques stress that AI tools often encode rationality/civility ideals through proxies that can privilege specific communicative styles and groups; this is a transparency problem as much as a bias problem, because the operationalization of norms is frequently opaque to participants (Carstens & Friess, 2024).

Design opportunities:

- Evidence-trace legibility: Representations of "clusters," "positions," or "insights" should expose a navigable link to supporting (and potentially refuting) human contributions, enabling users to validate whether the representation is warranted.
- Transformation ledger: Interfaces should communicate the mediation stages (e.g., preprocessing, clustering, summarization, scoring/selection) and their rationales as a compact, user-facing narrative, including the role of parameter settings and thresholds. This responds to the multi-stage nature of AI features in mass deliberation.
- Exclusions, scope, and uncertainty disclosures: Civic explainability requires making visible what is not represented and with what confidence representations are produced—especially for language-based inference and classification outputs, where uncertainty and hallucination risks are salient.
- Transparency of embedded deliberative norms: Where systems use proxies of deliberative quality (e.g., civility, argumentation structure), the interface should make these operational definitions explicit and open to scrutiny.

4.4.2 Gap B: Coherence pressures and the risk of epistemic closure

The corpus provides strong evidence that narrative and visualization can improve legibility for non-experts, yet it is less mature in specifying safeguards that preserve plurality when systems compress diverse contributions into coherent accounts. Narrative visualization research formalizes the spectrum between author-driven "telling" and reader-driven "exploring," but in civic settings the design tension intensifies: increasing coherence can inadvertently produce epistemic closure by narrowing what is visible, what is salient, and what is treated as "the" collective view (Jessica Hullman & Nick Diakopoulos, 2011; Segel & Heer, 2010). Empirical civic communication work shows that storytelling can increase citizen understanding relative to portals and dashboards and argues that usability and completeness may be more important than "neutral objectivity" for non-expert publics; however, this also implies that storytelling choices shape interpretation and thus require explicit accountability (Kempeneer et al., 2025). Related work in data storytelling and misinformation indicates that how text and visualization are integrated can materially influence interpretation and misinterpretation risks (Zheng & Ma, 2022), and that narrative techniques can have persuasive effects beyond comprehension (Chambers & Denham, 2025).

Democratic deliberation values disagreement as informational and normative input; plurality is not noise to be eliminated but a resource that must remain legible. If AI-mediated representations over-emphasize consensus, they may suppress minority voices or present contested issues

Chapter // 4

as settled, undermining legitimacy and contestability (Carstens & Friess, 2024; Tessler et al., 2024).

Design opportunities:

- Plural navigation instead of single-story summarization: Narrative/visual scaffolds should provide multiple, navigable lenses (e.g., per cluster, per stakeholder constellation, per contested axis), aligning with the author-reader balance described in narrative visualization theory.
- Disagreement-forward representations: Interfaces should intentionally surface tensions, counter-arguments, and minority clusters as first-class elements rather than as exceptions, particularly when algorithmic selection ranks "most agreed" content.
- Rhetorical transparency cues: Because narrative visualization frames interpretation, systems should communicate editorial choices (what is emphasized, what is omitted, why a framing was chosen) as part of civic accountability.
- Progressive disclosure under literacy constraints: Storytelling supports comprehension but does not automatically compensate for low data literacy; scaffolds should be staged and revisitable, enabling users to move from overview to details and evidence without losing orientation.

4.4.3 Gap C: Contestability remains a principle more than an interaction grammar

Contestability is widely acknowledged as necessary for legitimate AI mediation, but the corpus shows uneven translation into concrete, reusable interaction mechanisms that enable publics to challenge and steer representations. At the systems level, participatory approaches propose that stakeholders should be able to deliberate over ML pipeline choices (features, trade-offs, evaluation metrics), positioning models as boundary objects for negotiation rather than authoritative arbiters (Zhang et al., 2023). At the interface level, reflection nudges demonstrate that micro-interactions can shape deliberative quality upstream at the point of contribution (Yeo et al., 2024). Yet many tools still emphasize output

presentation over procedural affordances for contestation (Zeginis et al., 2026).

In civic settings, explainability should not only help users "understand" but also support legitimate challenge and revision when representations are disputed. Contestability is therefore both a design requirement and an evaluation target (Tessler et al., 2024; Tim Miller, 2019).

Design opportunities:

- Standardized interrogation pathways: An interaction repertoire links user questions to explanation types and actions. In fairness-aware planning, interactive "Why," "Why Not," "What If," and "How To" pathways support diagnosis and mitigation exploration, demonstrating how contestability can be enacted through action-oriented explanation.
- Structured critique and revision loops: AI-mediated deliberation protocols that include explicit critique phases operationalize contestability procedurally, not as an optional comment layer.
- Parameter and threshold legibility for re-aggregation: If clustering, scoring, or selection rules shape what becomes visible, users need controlled means to inspect and adjust these levers.
- Civic-scale navigation as agency: Tools that support filtering, clustering, and similarity-based exploration of large participation corpora can be reframed as agency mechanisms: they allow citizens to locate patterns, check redundancy, and connect contributions across scale.
- Upstream agency through contribution scaffolds: Reflection nudges exemplify how interface design can support deliberative agency before AI aggregation occurs.

4.4.4 Gap D: Evaluation remains trust-heavy and weakly aligned to democratic adequacy

The evaluation literature in the corpus repeatedly warns against treating "trust" as a sufficient outcome. Empirical and conceptual work distinguishes trust (attitude) from reliance (behavior), showing they can diverge; this undermines conclusions drawn from subjective trust measures

alone and motivates calibration-oriented evaluation (Rong et al., 2024; Scharowski et al., 2023). More broadly, XAI and HCXAI reviews emphasize heterogeneity in tasks, users, and metrics, challenging the portability of explanation "effects" across contexts (Barredo Arrieta et al., 2020; Ridley, 2025; Rong et al., 2024). In civic settings, additional dimensions become central: representativeness, minority inclusion, perceived fairness, and the procedural capacity to contest outcomes. Work on AI mediation evaluates not only endorsement and statement quality but also whether viewpoints (including minority voices) are represented and how agreement changes under mediation, illustrating legitimacy-oriented outcome dimensions (Tessler et al., 2024). Citizen-jury evidence further shows that explainability–accuracy trade-offs are context- and stakes-dependent, reinforcing the need for ecologically valid evaluation designs (van der Veer et al., 2021). Finally, narrative visualization research indicates that rhetorical and persuasive effects can be measurable and ethically consequential, implying that civic explainability evaluation should include potential framing side-effects (Chambers & Denham, 2025; Jessica Hullman & Nick Diakopoulos, 2011; Zheng & Ma, 2022).

Design opportunities:

- Multi-criterion evaluation model for civic explainability: Evaluate explanation interfaces against (i) comprehension/mental models, (ii) calibrated reliance under uncertainty, (iii) legitimacy and representational fairness (including minority inclusion), and (iv) contestability/recourse capacity.
- Role and literacy-sensitive protocols: Evidence shows that comprehension depends on data literacy and domain knowledge; civic evaluation must segment publics and roles rather than assuming a generic "user".
- Measure rhetorical side-effects: Given persuasive potential in narrative representations, evaluation should test for framing and persuasion effects alongside comprehension outcomes.
- Domain and stakes-dependent trade-off testing: Civic domains differ in acceptable trade-offs between explainability and accuracy, motivating context-specific evaluation rather than universal prescriptions.

4.4.5 From gaps to a communication-design focus

The SLR indicates multiple, mutually reinforcing gaps; addressing all of them at once would require infrastructural platform redesign and long-term field deployment. This thesis contribution was oriented through the analysis and feedback from the ORBIS European project, and the design opportunities taken in consideration as design requisites are aligned with it.

Two implications follow from the corpus. First, pipeline accountability emerges as a primary need: when AI systems cluster, summarize, and score deliberative contributions, the legitimacy of the resulting collective representations depends on evidential traceability, disclosure of exclusions and uncertainty, and transparency of norm-laden operationalizations (Bhatia & Sukthankar, 2025; Carstens & Friess, 2024; Zeginis et al., 2026). Second, interpretability at scale is not solved by explanation text alone: sensemaking requires visual and interactional scaffolds that preserve plurality while enabling users to interrogate patterns, tensions, and minority positions (Jessica Hullman & Nick Diakopoulos, 2011; Kempeneer et al., 2025; Segel & Heer, 2010). Contestability mechanisms, in turn, provide the procedural bridge between "seeing" and "challenging," and the evaluation literature motivates measuring civic adequacy in terms of comprehension, calibration, legitimacy, and contestability rather than trust alone (Lyu et al., 2024; Rong et al., 2024; Scharowski et al., 2023; Tessler et al., 2024).

5.0

FROM EVIDENCE TO DESIGN. A COMMUNICATION DESIGN FRAMEWORK FOR HCXAI

This chapter translates the SLR findings into an actionable communication design framework for HCXAI in civic deliberation. The framework responds to a recurrent limitation identified in the corpus: while XAI research offers a wide repertoire of explanation methods and taxonomies, it often under-specifies how those methods should be made publicly intelligible within complex socio-technical contexts, particularly when explanations must support heterogeneous audiences and downstream public reasoning (Barredo Arrieta et al., 2020; Herrera, 2025; Ridley, 2025). At the same time, narrative visualization and data storytelling research offers a complementary contribution that is essential for operationalization: it provides a design language for sequencing, attention guidance, annotation, and interaction that can scaffold comprehension under cognitive constraints, while also making visible the rhetorical risks of framing and persuasive effects in narrative representations (Chambers & Denham, 2025; Jessica Hullman & Nick Diakopoulos, 2011; Segel & Heer, 2010; Shao et al., 2024). Finally, civic-tech and deliberation research emphasizes that AI support is typically pipeline-shaped (moderation, structuring, summarization, aggregation) and that seemingly technical choices can encode normative assumptions and distributional effects, motivating plurality preservation and contestability as core requirements rather than optional features (Carstens & Friess, 2024; Zeginis et al., 2026).

Accordingly, the framework treats communication design as a mediating layer between computational mediation processes and civic reasoning practices. This mediating layer is enacted through three coupled instruments—narrative, visualization, and interaction—and is grounded in four synthesis dimensions from Chapter 4: transformation transparency, plurality-preserving sensemaking, agency/contestability mechanisms and evaluation beyond trust. The storytelling lens sharpens the operational consequences of these dimensions: it foregrounds that (i) sequencing is part of explainability (how the pipeline is narrated and staged), (ii) evidence binding is a civic safeguard (claims must be visibly tied to sources), and (iii) plurality can be preserved through modular narratives rather than a single authoritative storyline (Borges et al., 2025; Jessica Hullman & Nick Diakopoulos, 2011; Kempeneer et al., 2025; Segel & Heer, 2010; Zheng & Ma, 2022).

5.1 Translating Review Findings into Design Requirements

To move from evidence to design, this chapter operationalizes the synthesis from Chapter 4 into a set of conceptual design requirements for HCXAI in civic deliberation. Here, “requirements” are intended in a system-agnostic sense: they define the communicative and interactional conditions that an AI-supported deliberation system should satisfy to remain intelligible, contestable, and civically adequate for heterogeneous publics. They do not prescribe specific interface features or technical solutions; rather, they express what must be supported—across narrative framing, visual encoding, and interaction design—so that downstream design choices can be evaluated against explicit criteria.

The requirements were derived by treating the gaps and design opportunities identified in Chapter 4 as generative directions rather than as a one-to-one translation template. In other words, design opportunities informed the requirements as directional cues: a single opportunity may yield multiple requirements when it implies distinct communicative constraints, and multiple opportunities may converge into a single requirement when they point to a shared design condition. This avoids an artificial “mapping” that would overfit the table to the wording of Chapter 4, while preserving traceability to the underlying evidence.

Each requirement is grounded in the SLR corpus through an evidence field that lists the paper IDs from the extraction matrix in which the requirement is supported or motivated. This evidence anchoring serves two functions. First, it makes the framework auditable: readers can trace each requirement back to the empirical or conceptual claims from the reviewed literature. Second, it makes explicit where requirements rest on convergent evidence across subfields (XAI, civic-tech, narrative visualization) versus where they rely on fewer or more interpretive sources—an important consideration given the socio-technical and normative stakes of civic deliberation systems.

Table 8 presents the resulting conceptual requirements. Each row includes (i) a requirement ID, (ii) a concise label, (iii) a descriptor specifying the intended design condition, and (iv) the paper IDs that provide supporting evidence in the extraction matrix. At this stage, the requirements are intentionally kept granular and are not merged or collapsed, to preserve analytical resolution before the next step: comparison with the ORBIS project requirements.

Tab. 08 Design requirements emerged from the literature analysis

Req. ID	Requirement	Descriptor	Evidence
R01	Progressive Disclosure & Narrative Sequencing	Designers must structure complex information using narrative sequencing (e.g., “scrollytelling,” “Martini Glass” structure). Present context and conceptual framing before technical details or visualizations to scaffold understanding and prevent cognitive overload.	P01, P07, P20, P26
R02	Audience-Adaptive Complexity	The complexity of visualizations and explanations must be adapted to the target audience’s domain knowledge and literacy levels. Designers should account for different epistemic profiles (e.g., experts vs. lay citizens) and education levels, which moderate persuasion and comprehension.	P01, P06, P13, P17, P27
R03	Cross-Modal Evidential Anchoring	Textual claims (narratives, summaries, AI insights) must be explicitly linked to visual evidence. Use interactive mechanisms like hover-to-highlight or hyperlinks to bind high-level insights back to specific source data (e.g., original citizen proposals/comments) to support verification and reduce text dominance.	P04, P05, P14, P15, P16
R04	Explicit Provenance & Authorship Marking	The interface must visually distinguish between human-authored content and AI-generated syntheses (e.g., using specific icons or distinct visual styles). Designers must employ provenance rhetoric (citations, methodology notes) to maintain credibility and allow users to assess the source of the information.	P02, P06, P14, P24
R05	“Seamful” Design & Uncertainty Communication	Avoid presenting AI outputs or data stories as flawless or neutral truths. Interfaces should practice seamful design by visibly exposing system limitations, uncertainties, and data gaps (e.g., model reality vs. lived reality) to encourage epistemic vigilance and prevent over-reliance.	P05, P18, P20, P22, P25
R06	Upstream Reflection Scaffolds	Incorporate reflection nudges (e.g., persona prompts, storytelling prompts) during the drafting/input phase. These communicative scaffolds intervene upstream to improve the quality, diversity, and constructiveness of user contributions before they enter the public discussion.	P11, P12
R07	Interactive “What-If” & Comparison Tools	Move beyond static explanations by providing what-if simulations, comparative views (contrastive explanations), and filtering capabilities. This supports human-centered reasoning by allowing users to test hypotheses, see consequences of changes, and compare diverse scenarios.	P08, P19, P26, P28
R08	Legible & Familiar Visualization Forms	Prioritize domain-legible and accessible visualization types (e.g., simple histograms, heatmaps, traffic-light indicators) over complex forms (e.g., large network graphs) that may intimidate non-expert participants. Ensure visual metaphors align with users’ mental models.	P16, P19, P20
R09	Support for Traceability, Contestability & History	The system must allow users to trace how AI outputs were produced (from source data to clustering/summaries) and to navigate the history of the discussion. Provide interaction primitives such as Extract (save/share evidence) and Replay (undo/view history) so users can contest interpretations, reproduce views, and revisit earlier states.	P17, P21, P22, P26
R10	Representation of Plurality & Conflict	AI synthesis tools should not artificially smooth over disagreements to create a false consensus. Design representations (e.g., clusters, argument maps) that make minority viewpoints, distinct issue communities, and conflict structures visible and legitimate.	P09, P10, P14, P15, P21
R11	Defensive Design Against Manipulation	Anticipate that narrative and rhetorical choices (omission, emphasis, framing) can manipulate interpretation. Implement defensive design strategies (e.g., exposing alternative framings, rigorous fact-checking layers) to protect against misinformation and dark patterns in data storytelling.	P04, P06, P22, P24, P25
R12	Facilitation of “Common Ground” via Mediation	Use AI as a mediator to draft synthesized statements that represent the group’s common ground. Design a workflow that allows participants to critique and revise these AI drafts, ensuring the final narrative is a negotiated consensus rather than an algorithmic imposition.	P12, P21
R13	End-to-End Transformation Ledger	The interface must expose the main transformations applied to deliberative input (selection, filtering, clustering, summarization, ranking), including what is excluded and why. Provide an end-to-end transformation ledger that supports accountability and interpretation of AI-mediated outputs.	P05, P12, P17, P22, P26
R14	Disclosure of Normative Operationalizations	When the system operationalizes normative concepts (e.g., rationality, civility, participation quality, fairness), it must disclose definitions, features/labels, and boundary cases with concrete examples. The operationalization itself should be inspectable and open to user critique.	P09, P10, P12, P17, P22

R15	Legibility of Parameters, Thresholds & Re-Aggregation Controls	Users should be able to inspect the key parameters, thresholds, and metrics that drive aggregation and visibility (e.g., cluster granularity, salience scoring, filtering rules). Provide robust defaults plus safe controls or sensitivity views that reveal how outcomes change under different settings.	P05, P10, P19, P26
R16	Editorial & Rhetorical Transparency Cues	Make editorial choices legible (ordering, emphasis, omissions, visual rhetoric) and surface how these choices shape interpretation. Provide alternative framings or counter-views to reduce the risk of persuasive bias being mistaken for neutrality.	P04, P06, P20, P24, P25
R17	In-Situ Feedback, Disagreement Capture & Revision Workflow	Provide mechanisms for participants to challenge AI outputs (clusters, summaries, labels) with structured feedback (e.g., wrong grouping, missing viewpoint, distorted summary). Preserve an audit trail of revisions and rationales, and support critique-driven iteration on mediated statements.	P12, P17, P21, P22, P26

The next section will use the ORBIS requirements previously introduced as relevant in Chapter 3 as a feasibility and scope filter. The outcome will be a reduced set of requirements that are both literature-grounded and project-relevant, which will then be reformulated into implementable design principles for the platform presented in Chapter 6.

5.1.1 Requirement catalogue

The following sections take the requisites that emerged from the literature gaps and potential design opportunities and compare them to the requirements from ORBIS filtered in chapter 3 based on the relevance to the Knowledge Graph visualization platform that will be introduced in chapter 6. This filtering process is meant to align the general requirements from the systematic literature review and ground them in a real case scenario.

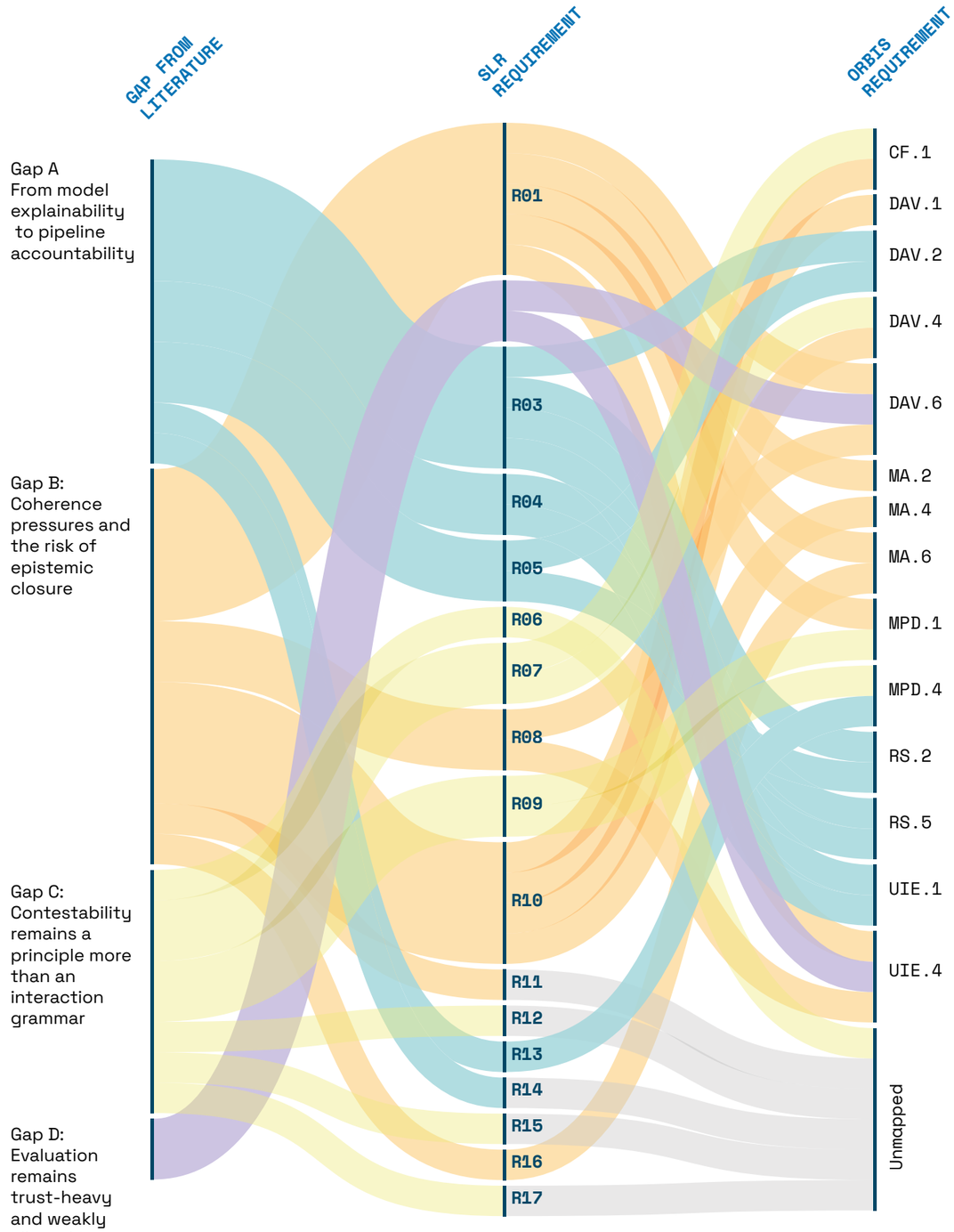


Fig.11 SLR requirements mapped to literature Gaps and ORBIS requisites

Gap from literature (Ch.4)	Design opportunity (Ch.4)	SLR Requirement ID	Requirement	Descriptor	Evidence	Link to ORBIS requirement (IDs)
Gap B: Coherence pressures and the risk of epistemic closure	Progressive disclosure under literacy constraints	R01	Progressive Disclosure & Narrative Sequencing	Designers must structure complex information using narrative sequencing (e.g., "scrollytelling," "Martini Glass" structure). Present context and conceptual framing before technical details or visualizations to scaffold understanding and prevent cognitive overload.	P01, P07, P20, P26	DAV.6, MA.2, MA.6, MPD.1, UIE.4
Gap D: Evaluation remains trust-heavy and weakly aligned to democratic adequacy	Role and literacy-sensitive protocols	R02	Audience-Adaptive Complexity	The complexity of visualizations and explanations must be adapted to the target audience's domain knowledge and literacy levels. Designers should account for different epistemic profiles (e.g., experts vs. lay citizens) and education levels, which moderate persuasion and comprehension.	P01, P06, P13, P17, P27	DAV.6, UIE.4
Gap A: From model explainability to pipeline accountability	Evidence-trace legibility	R03	Cross-Modal Evidential Anchoring	Textual claims (narratives, summaries, AI insights) must be explicitly linked to visual evidence. Use interactive mechanisms like hover-to-highlight or hyperlinks to bind high-level insights back to specific source data (e.g., original citizen proposals/comments) to support verification and reduce text dominance.	P04, P05, P14, P15, P16	DAV.2, RS.2, RS.5, UIE.1
Gap A: From model explainability to pipeline accountability	—	R04	Explicit Provenance & Authorship Marking	The interface must visually distinguish between human-authored content and AI-generated syntheses (e.g., using specific icons or distinct visual styles). Designers must employ provenance rhetoric (citations, methodology notes) to maintain credibility and allow users to assess the source of the information.	P02, P06, P14, P24	"RS.5, UIE.1"
Gap A: From model explainability to pipeline accountability	Exclusions, scope, and uncertainty disclosures	R05	"Seamful" Design & Uncertainty Communication	Avoid presenting AI outputs or data stories as flawless or neutral truths. Interfaces should practice seamful design by visibly exposing system limitations, uncertainties, and data gaps (e.g., model reality vs. lived reality) to encourage epistemic vigilance and prevent over-reliance.	P05, P18, P20, P22, P25	DAV.2, RS.2
Gap C: Contestability remains a principle more than an interaction grammar	—	R06	Upstream Reflection Scaffolds	Incorporate reflection nudges (e.g., persona prompts, storytelling prompts) during the drafting/input phase. These communicative scaffolds intervene upstream to improve the quality, diversity, and constructiveness of user contributions before they enter the public discussion.	P11, P12	—
Gap C: Contestability remains a principle more than an interaction grammar	Standardized interrogation pathways	R07	Interactive "What-If" & Comparison Tools	Move beyond static explanations by providing what-if simulations, comparative views (contrastive explanations), and filtering capabilities. This supports human-centered reasoning by allowing users to test hypotheses, see consequences of changes, and compare diverse scenarios.	P08, P19, P26, P28	CF.1, DAV.4
Gap B: Coherence pressures and the risk of epistemic closure	—	R08	Legible & Familiar Visualization Forms	Prioritize domain-legible and accessible visualization types over complex forms that may intimidate non-expert participants. Ensure visual metaphors align with users' mental models.	P16, P19, P20	DAV.6, UIE.4
Gap C: Contestability remains a principle more than an interaction grammar	—	R09	Support for Traceability, Contestability & History	The system must allow users to trace how AI outputs were produced (from source data to clustering/summaries) and to navigate the history of the discussion. Provide interaction primitives such as Extract (save/share evidence) and Replay (undo/view history) so users can contest interpretations, reproduce views, and revisit earlier states.	P17, P21, P22, P26	MPD.1, MPD.4
Gap B: Coherence pressures and the risk of epistemic closure	Disagreement-forward representations	R10	Representation of Plurality & Conflict	AI synthesis tools should not artificially smooth over disagreements to create a false consensus. Design representations (e.g., clusters, argument maps) that make minority viewpoints, distinct issue communities, and conflict structures visible and legitimate.	P09, P10, P14, P15, P21	CF.1, DAV.1, DAV.4, MA.4
Gap B: Coherence pressures and the risk of epistemic closure	—	R11	Defensive Design Against Manipulation	Anticipate that narrative and rhetorical choices (omission, emphasis, framing) can manipulate interpretation. Implement defensive design strategies (e.g., exposing alternative framings, rigorous fact-checking layers) to protect against misinformation and dark patterns in data storytelling.	P04, P06, P22, P24, P25	—

Gap from literature (Ch.4)	Design opportunity (Ch.4)	SLR Requirement ID	Requirement	Descriptor	Evidence	Link to ORBIS requirement (IDs)
Gap C: Contestability remains a principle more than an interaction grammar	Structured critique and revision loops	R12	Facilitation of "Common Ground" via Mediation	Use AI as a mediator to draft synthesized statements that represent the group's common ground. Design a workflow that allows participants to critique and revise these AI drafts, ensuring the final narrative is a negotiated consensus rather than an algorithmic imposition.	P12, P21	—
Gap A: From model explainability to pipeline accountability	Transformation ledger	R13	End-to-End Transformation Ledger	The interface must expose the main transformations applied to deliberative input (selection, filtering, clustering, summarization, ranking), including what is excluded and why. Provide an end-to-end transformation ledger that supports accountability and interpretation of AI-mediated outputs.	P05, P12, P17, P22, P26	MPD.4
Gap A: From model explainability to pipeline accountability	Transparency of embedded deliberative norms	R14	Disclosure of Normative Operationalizations	When the system operationalizes normative concepts (e.g., rationality, civility, participation quality, fairness), it must disclose definitions, features/labels, and boundary cases with concrete examples. The operationalization itself should be inspectable and open to user critique.	P09, P10, P12, P17, P22	—
Gap C: Contestability remains a principle more than an interaction grammar	Parameter and threshold legibility for re-aggregation	R15	Legibility of Parameters, Thresholds & Re-Aggregation Controls	Users should be able to inspect the key parameters, thresholds, and metrics that drive aggregation and visibility (e.g., cluster granularity, salience scoring, filtering rules). Provide robust defaults plus safe controls or sensitivity views that reveal how outcomes change under different settings.	P05, P10, P19, P26	—
Gap B: Coherence pressures and the risk of epistemic closure	Rhetorical transparency cues	R16	Editorial & Rhetorical Transparency Cues	Make editorial choices legible (ordering, emphasis, omissions, visual rhetoric) and surface how these choices shape interpretation. Provide alternative framings or counter-views to reduce the risk of persuasive bias being mistaken for neutrality.	P04, P06, P20, P24, P25	MA.6
Gap C: Contestability remains a principle more than an interaction grammar	Structured critique and revision loops	R17	In-Situ Feedback, Disagreement Capture & Revision Workflow	Provide mechanisms for participants to challenge AI outputs (clusters, summaries, labels) with structured feedback (e.g., wrong grouping, missing viewpoint, distorted summary). Preserve an audit trail of revisions and rationales, and support critique-driven iteration on mediated statements.	P12, P17, P21, P22, P26	—

Tab. 09 SLR requisites mapped to the literature Gaps, the design opportunities and the ORBIS requirements filtered

5.2 Design Principles for HCXAI in Civic Deliberation

Building on the conceptual requirements derived from the SLR, this section formulates a set of design principles that translate “what must be supported” into actionable directions for interface and interaction design in AI-supported civic deliberation. While the requirements remain system-agnostic, the principles are explicitly oriented toward implementation and evaluation: each principle specifies a design intent that can be operationalized in concrete interface mechanisms and later assessed against civic explainability goals.

The derivation follows two steps. First, the requirements were screened through project relevance by considering whether they have correspondence with ORBIS requirements (as reported in the “Link to ORBIS requirement (IDs)” column). Second, the ORBIS-linked requirements were recomposed into a smaller set of principles by grouping those that jointly support the same communicative function (e.g., comprehension scaffolding, accountability, plurality-preserving sensemaking). This does not invalidate requirements without ORBIS links; rather, it identifies which requirements can be credibly pursued within the scope of the platform presented in Chapter 6, while keeping the remaining ones as literature-grounded criteria for critical reflection and future implementations.

DP1 - Stage civic explainability through progressive disclosure

AI-mediated deliberation should be communicated as a staged experience that scaffolds understanding: begin with orienting frames (what the discussion is about, what the system does), then introduce analytical constructs (clusters, viewpoints, summaries), and only then offer deeper technical detail and advanced exploration. This principle treats sequencing as part of explainability, reducing cognitive overload and supporting heterogeneous literacy levels.

Derived from: R01, R02 (and reinforced by R08 where complex views are used).
ORBIS links: DAV.6, MA.2, MA.6, MPD.1, UIE.4.

DP2 - Prefer legible visual languages, especially for graph-based exploration

Where complex visual forms (e.g., network graphs) are necessary, they should be made publicly legible through familiar encodings, clear interaction cues, and consistent visual metaphors aligned with users’ mental models. This includes reducing visual complexity by default and providing guided entry points for non-experts, without preventing expert-level exploration.

Derived from: R08 (supported by R01/R02 as enabling conditions).
ORBIS links: DAV.6, UIE.4.

DP3 - Bind every synthesized claim to inspectable evidence and explicit provenance

Narratives, summaries, clusters, and analytic insights must remain verifiable in the interface. Users should be able to trace high-level claims back to concrete sources (original contributions, excerpts, or linked evidence) through direct interaction (e.g., highlight-on-hover, “show sources”, citation-style linking). In parallel, the system should explicitly mark authorship and provenance (human vs AI; synthesis vs original), so that users can correctly interpret status and responsibility.

Derived from: R03, R04.
ORBIS links: DAV.2, RS.2, RS.5, UIE.1.

DP4 - Make mediation “seams” visible: uncertainty, exclusions, and limits are first-class content

Because deliberative outputs are produced through aggregation and abstraction, the interface should not present AI-generated representations as neutral facts. It must disclose uncertainty, incompleteness, and known limitations (e.g., coverage gaps, ambiguous clusters, low-confidence summaries), and it should make exclusions and scope boundaries legible to users. This enables calibrated reliance and reduces the risk of overstating civic authority.

Derived from: R05.
ORBIS links: DAV.2, RS.2.

DP5 - Represent plurality and conflict rather than collapsing disagreement into coherence

Design should preserve the structure of disagreement: clusters and viewpoint representations should make distinct issue communities visible, including minority or less frequent perspectives. Summaries should avoid producing a single authoritative storyline when the underlying discourse is plural, and comparison views should support seeing overlaps and disjoints across positions without implying forced convergence.

Derived from: R10.
ORBIS links: CF.1, DAV.1, DAV.4, MA.4.

DP6 - Provide structured comparison and interrogation pathways

Users should be able to interrogate the deliberative space through standardized interaction pathways: comparing viewpoints, filtering and regrouping content, exploring overlaps/disjoints, and testing alternative readings of the discourse. The goal is to shift from passive consumption of AI outputs to active sensemaking and hypothesis testing by citizens and facilitators.

Derived from: R07 (and practically intertwined with R10 for plural discourse).
ORBIS links: CF.1, DAV.4.

DP7 - Ensure traceability across time and phases of deliberation

Civic deliberation is dynamic and often multi-phase. The interface should support temporal navigation (what changed, when, and why) and provide traceability across phases, including the ability to revisit earlier states of the discussion and inspect how outputs evolved. This supports contestability, learning, and institutional accountability.

Derived from: R09.
ORBIS links: MPD.1, MPD.4.

DP8 - Expose the transformation pipeline as an interpretable ledger

Beyond local traceability, the system should make the overall transformation chain interpretable: how raw contributions are

selected, filtered, clustered, summarized, and ranked into public-facing representations. A transformation ledger supports accountability by clarifying what operations shaped the output and where normative or technical choices may have influenced what becomes visible.

Derived from: R13 (complementary to R09 and R05).
ORBIS links: MPD.4.

DP9 - Treat narrative and infographic outputs as rhetorical acts that require transparency

When the system produces narrative devices (e.g., event evolution infographics), it should make editorial choices explicit: what is emphasized or omitted, how ordering shapes interpretation, and what alternative framings could exist. This principle recognizes that narrative representation is never neutral and that rhetorical transparency is necessary to prevent persuasion effects from being mistaken for objective explanation.

Derived from: R16 (often coordinated with R01).
ORBIS links: MA.6.

5.3 Communication Design as a Mediating Layer in AI-Supported Deliberation

In AI-supported civic deliberation, communication design mediates between two domains that operate with different logics and legitimacy conditions: on the one hand, computational mediation processes that transform heterogeneous contributions into collective representations; on the other hand, civic reasoning practices through which publics interpret, contest, and act on those representations. The core claim of this chapter is that public explainability is achieved (or fails) at this interface, where meaning is negotiated through representational choices, narrative framing, and interaction pathways rather than through model-centric justifications alone.

5.3.1 Mediation as a communicative contract for civic explainability

Treating communication design as a mediating layer reframes explainability as a communicative contract: the system makes an implicit promise about what its representations mean, what they can support, and what users are entitled to question. In deliberative contexts, that contract

must satisfy civic constraints that go beyond usability: users need to understand how collective outputs were produced, what was excluded or simplified, how disagreement is represented, and how their agency is preserved when AI systems structure the public record. This is why the design principles in Section 5.2 emphasize mechanisms that sustain comprehension, verification, and contestation under conditions of plurality and asymmetrical expertise.

5.3.2 Three coupled instruments: narrative, visualization, interaction

The mediating layer is enacted through three coupled instruments—narrative, visualization, and interaction—that function together as a civic explainability apparatus:

Narrative provides sequencing and interpretive framing. In this framework, narrative is not synonymous with "telling a story" in an editorial sense; it is the mechanism through which a pipeline is staged into intelligible steps and through which users are oriented to what the system is doing and why it matters. This is operationalized in DP1 (progressive disclosure) and extended in DP9 (rhetorical transparency for narrative/infographic outputs), recognizing that narrative devices carry rhetorical force and therefore require explicit accountability cues.

Visualization provides perceptual structure for plurality and evidence. In deliberation, the challenge is not only to show results, but to represent difference, conflict, and distribution without collapsing them into a single authoritative view. Principles DP2 (legible visual languages) and DP5 (plurality and conflict) treat visualization as a civic instrument: it shapes what becomes salient, what becomes comparable, and what remains visible as legitimate disagreement.

Interaction provides agency, interrogation, and procedural contestability. Interaction mechanisms determine whether AI-mediated representations are only consumed passively or can be questioned, decomposed, compared, and revisited. DP6 (structured comparison and interrogation pathways), DP7 (traceability across time and phases), and DP8 (transformation ledger) frame contestability as a procedural capability rather than an abstract principle.

These instruments are interdependent: narrative without interaction risks interpretive closure; visualization without evidence binding risks over-authoritative readings; interaction without legible framing risks becoming expert-only and excluding non-specialist publics.

5.3.3 How the principles operationalize the four synthesis dimensions

Section 5.2 can be read as a design-level operationalization of the four synthesis dimensions articulated at the beginning of this chapter.

Transformation transparency.
This dimension is implemented by a combination of evidence binding and pipeline intelligibility. DP3 ensures that synthesized claims remain tethered to inspectable sources and explicit provenance, while DP8 makes the overall chain of transformations interpretable as a ledger rather than a black-box mediation. DP4 complements both by making uncertainty, exclusions, and limits visible, preventing outputs from being misread as exhaustive civic facts. Together, these principles treat transparency as a communicative achievement: not "more information," but structured legibility of what changed, why, and with what confidence.

Plurality-preserving sensemaking.
Plurality is not a visualization style; it is a commitment to represent disagreement without forcing coherence. DP5 specifies the representational stance (conflict is preserved rather than smoothed), and DP6 specifies the user pathways needed to navigate that plurality (comparison, filtering, overlap/disjoint exploration). Here, communication design functions as a safeguard against epistemic closure: it supports understanding while resisting the tendency of narratives and summaries to collapse contested spaces into single storylines.

Agency and contestability mechanisms.
Agency is preserved when users can interrogate, revisit, and contextualize system-mediated representations. DP6 provides structured interrogation pathways; DP7 extends agency across time and phases, making deliberation navigable as a process rather than a static snapshot; DP8 provides the accountability substrate that makes contestation intelligible. This turns contestability into a usable civic capacity: the interface does not explain, it supports procedures of challenge, verification, and re-reading.

Evaluation beyond trust.
Although trust is not foregrounded as a standalone principle in the ORBIS-aligned set, the framework implicitly targets outcomes that exceed trust self-reports: comprehension support (DP1/DP2), verifiability (DP3), calibrated interpretation under uncertainty (DP4), legitimacy through plural representation (DP5), and procedural accountability (DP7/DP8). The mediating layer provides evaluation handles: it produces observable

interaction behaviors (evidence inspection, comparison, temporal revisiting) that can be assessed as proxies for civic adequacy rather than as acceptance.

5.3.4 Rhetorical accountability: why “how it is shown” is part of what it means

A key implication of treating communication design as mediation is that representational choices become normatively consequential. Ordering, emphasis, annotation, and aggregation are not neutral; they shape interpretation and can redistribute visibility across positions. This is why DP9 frames narrative and infographic outputs as rhetorical acts that require transparency: the system must clarify what is emphasized or omitted and make room for alternative framings when representing contested civic material. The goal is not to eliminate rhetoric—impossible in any communicative artifact—but to render rhetorical influence legible and accountable as part of the explainability contract.

5.3.5 Implications for Chapter 6: from principles to platform mechanisms

The framework developed in Chapter 5 positions the platform in Chapter 6 as an instantiation of a broader design stance: AI-mediated deliberation becomes publicly intelligible when its mediation is communicated through staged onboarding, legible plural representations, evidence-anchored synthesis, and accountable interaction pathways. Chapter 6 will therefore not be evaluated only by whether it “shows results,” but by whether it implements the mediating layer implied by the principles: making transformations inspectable, preserving plural civic meaning, and enabling citizens and facilitators to move from consumption to inquiry.

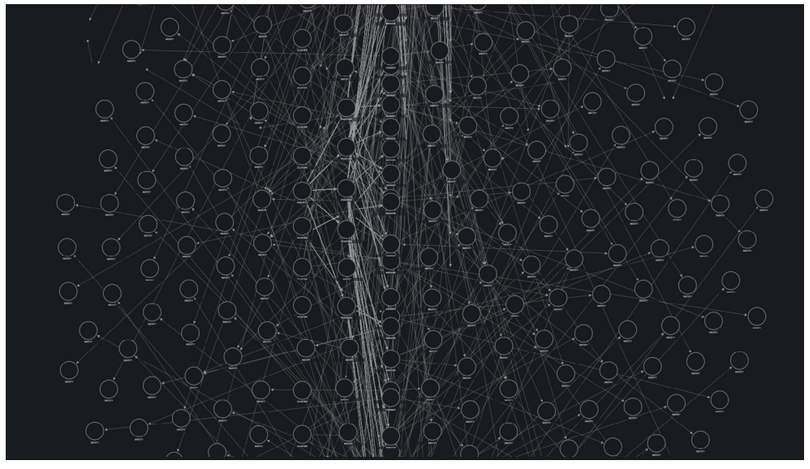
[p_136]

6.0
INTERACTIVE KNOWLEDGE GRAPH. A COMMUNICATION APPROACH FOR MAPPING ORBIS DELIBERATIONS

ORBIS adopts advanced visualisation analytics to make large-scale deliberation intelligible, navigable, and actionable for heterogeneous audiences, under the premise that linear reading breaks down when contributions and interconnections scale beyond what users can reconstruct manually. Within this ecosystem, the Knowledge Graph (KG) functions as a core representational layer: it models deliberative discourse as a network of interconnected positions, arguments, clusters, and extracted concepts, enabling users to explore how viewpoints relate, where agreement/disagreement accumulates, and which concepts bridge otherwise separate parts of the debate.

However, the first KG view available through the ORBIS integration—rendered in Grafana—was primarily a technical output rather than a communicative instrument. Its nodes and links appeared visually homogeneous, making deliberative roles (e.g., position vs argument vs extracted entity) indistinguishable at a glance. As a consequence, interpretation depended on repeated clicking and panel inspection, raising interaction costs and fragmenting the reading experience precisely when users need stable context to reason about a dense debate. Figure 12 documents this baseline and provides the motivation for a design shift: from “showing connectivity” to communicating the structure of deliberation.

Fig.12 Knowledge Graph starting point (Grafana)



This chapter presents Il Corollario as a communication-design intervention on the ORBIS KG: an interactive interface that operationalises the design principles established in Chapter 5 through concrete data structures, encodings, and interaction pathways. In line with ORBIS’ framing of the KG as a human-centred explainability device—where AI surfaces patterns while

judgement remains with human actors—the design goal is not to automate interpretation, but to scaffold it through legible representational cues, progressive disclosure of complexity, and evidence-aware inspection of nodes and relations. The sections that follow unpack this operationalisation from the inside out: (i) the data model and KG structure; (ii) the visual grammar and interaction design; (iii) the sensemaking affordances enabled by network exploration; and (iv) the integration of the KG as a reusable reporting and analysis component within the ORBIS ecosystem.

6.1 Data Model and Knowledge Graph Structure

In ORBIS, deliberative content is harvested from third-party platforms and processed through an event-driven pipeline (storage in MinIO, orchestration via Kafka, enrichment and analytics modules, and delivery through the ORBIS API). For the KG specifically, ORBIS generates sub-topologies scoped to a single discussion topic and exposes each of them through a dedicated URL that can be embedded or referenced from the host deliberation platform, allowing users to explore the relevant KG segment without leaving the deliberation environment.

This delivery logic matters for communication design because it establishes a stable unit of interpretation: the KG is not presented as an abstract global knowledge base, but as an interpretable map of a bounded deliberative event (one topic, one debate). This boundedness supports calibrated reading: users can treat graph patterns (hubs, bridges, isolated nodes) as meaningful within the debate’s scope, rather than as decontextualised social signals.

Trade-off: Topic scoping improves interpretability and performance, but it also constrains cross-topic comparisons (e.g., recurring themes across multiple debates). In Il Corollario, this limitation is accepted as a deliberate focus on within-topic sensemaking, with cross-topic synthesis left to other ORBIS reporting components.

6.1.1 A typed, deliberation-oriented schema

ORBIS frames KG construction as triplet-based (node–relationship–node) and supports it with an ontological schema that standardises how deliberation elements connect. Il Corollario adopts this principle of explicit structure and translates it into a typed argumentation-centric graph that prioritises deliberative legibility over generic graph completeness.

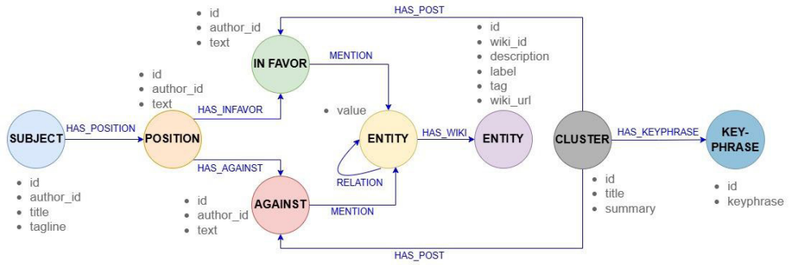


Fig.13 ORBIS knowledge graph data structure

At the level of the interactive visualisation, the schema is intentionally heterogeneous: different node types represent different deliberative functions (topic framing, viewpoints, supporting/opposing contributions, thematic aggregation, and semantic annotation). This decision directly supports the platform's interpretive contract: the graph should not only state that two items are linked, but also communicate what kind of deliberative relation the link represents.

Node types implemented in Il Corollario:

- SUBJECT: the discussion root (prompt / topic question).
- POSITION: main viewpoints (the debate's stance options).
- INFAVOR / AGAINST: stance-marked argument contributions connected to a position.
- CLUSTER: AI-generated thematic group nodes that connect to their member arguments.
- ENTITY: AI-extracted keywords used as semantic bridges between arguments (semantic annotation layer).

This separation is consistent with ORBIS' KG redesign rationale: it distinguishes debate structure (subject/positions/arguments) from conceptual annotation (keywords/entities), preventing semantic extraction from being mistaken as equivalent in status to human-authored contributions.

Edge predicates implemented in Il Corollario (prototype dataset):

- HAS_POSITION (SUBJECT → POSITION)

- HAS_INFAVOR / HAS_AGAINST (POSITION → argument)
- HAS_POST (CLUSTER → argument)
- MENTION (argument → ENTITY keyword)

In Chapter 5 terms, this schema-level separation is the first step in turning a "black-box" analytic output into an inspectable civic representation: users can recognise which elements originate in participation (positions/arguments) and which are produced through computational mediation (clusters/keywords). The expected value is not only comprehension but procedural interpretability: the user can decide what to treat as evidence, what to treat as annotation, and what to treat as a hypothesis generated by the system.

6.1.2 Provenance-aware node payloads

A core design move in Il Corollario is that node records carry enough payload to support details-on-demand without severing ties to source content. In ORBIS' broader KG architecture, deliberation items are stored alongside metadata and are expected to be retrievable as structured inputs/outputs (CSV/JSON/API streams), enabling downstream visual representations.

In the Il Corollario implementation, this is operationalised through a lightweight export format:

- KG_nodes.csv stores the node identifier plus a nested "detail" payload (e.g., author ID, contribution text, short summary/tagline, platform-level IDs).
- KG_edges.csv stores source–target relationships plus an explicit predicate label (mainStat).

This "payload + predicate" approach is a pragmatic compromise between two needs:

- Graph readability (keep the visualisation responsive by avoiding heavyweight DB queries at runtime).
- Auditability (retain the minimum provenance necessary for users to verify what a node represents when inspecting it).

Trade-off: CSV-serialised payloads are well suited for a standalone prototype and reproducible demos, but they are not a substitute for live API-backed provenance in production deployments. In a full ORBIS integration, the same fields would ideally resolve to authenticated endpoints (e.g., opening the original contribution on the host platform), consistent with ORBIS' API-oriented architecture.

6.1.3 Data hygiene as a communicative requirement

A subtle but consequential issue in KG visualisation is that redundant relations can distort perceived salience: multiple identical edges effectively "thicken" the visual connection and can be misread as stronger evidence of association. ORBIS explicitly addresses this in the KG redesign by removing duplicate edges during ingestion (unique key on source–target–relation) to prevent redundant visual weight.

Il Corollario implements the same logic at load time: before rendering, edge lists are deduplicated so that connectivity reflects structural relations rather than artefacts of export or processing repetition. The expected user value is interpretive fairness: the interface should not amplify a connection simply because it appears multiple times in the raw export.

6.1.4 Summary: why the data model is part of the explainability design

In Il Corollario, "explainability" begins before visual encoding: it is embedded in the data model through (i) typed deliberation roles, (ii) explicit relation predicates, and (iii) provenance-carrying payloads that support inspection. This is consistent with ORBIS' positioning of the KG as a human-centred explainability instrument, where AI-produced structure remains navigable and tethered to deliberative content rather than presented as an opaque analytic verdict.

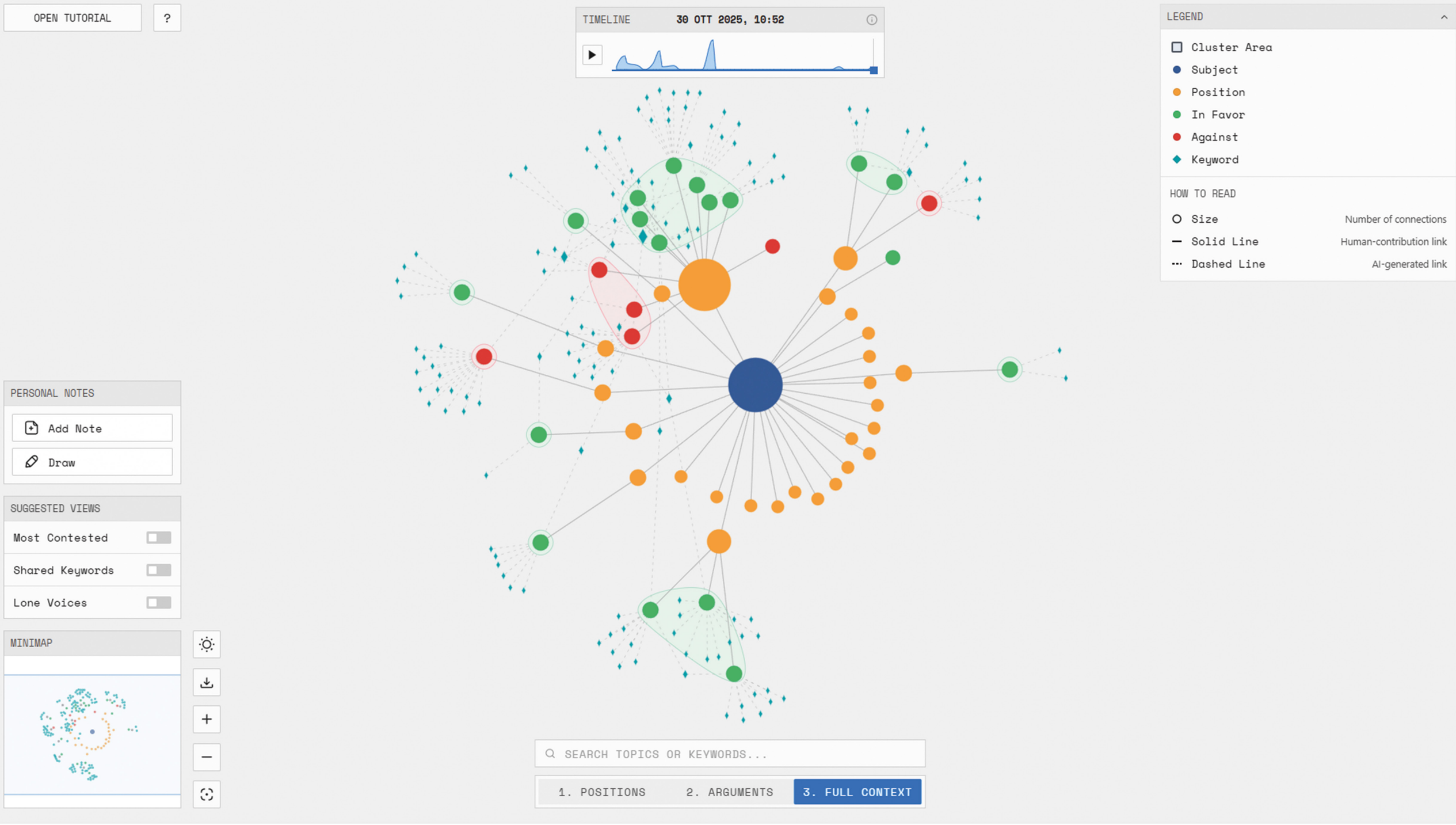
6.2 Visual Encoding and Interaction Design

This section explains how Il Corollario turns the ORBIS KG from a "connectivity display" into a readable deliberation map by implementing an explicit visual grammar and a tightly coupled interaction grammar. The design goal is to make users perceive what each node represents (and what kind of relation each edge expresses) before they click—reducing the "panel-driven" exploration cost documented in the baseline Grafana KG.

This operationalises the Chapter 5 principles on legible visual languages (DP2), plurality-preserving representation (DP5), and rhetorical transparency around salience cues (DP9). It also targets ORBIS requirements for user-friendly interfaces for network graphs and viewpoints (UIE.4, DAV.6).

IlCorollario

Fig.14 Il Corollario full-screen visualization



Tutorial &
From data to Knwoledge

Timeline

Legend Panel

Fig.15 Il Corollario full-screen components indicators

OPEN TUTORIAL ?

TIMELINE 30 OTT 2025, 10:52

LEGEND

- Cluster Area
- Subject
- Position
- In Favor
- Against
- Keyword

HOW TO READ

Size	Number of connections
Solid Line	Human-contribution link
Dashed Line	AI-generated link

Personal notes
& drawing panel

PERSONAL NOTES

Add Note

Draw

Suggested views
panel

SUGGESTED VIEWS

Most Contested

Shared Keywords

Lone Voices

MINIMAP

SEARCH TOPICS OR KEYWORDS...

1. POSITIONS 2. ARGUMENTS 3. FULL CONTEXT

Navigation
Panel

Depth-filters
& Searchbar

Details panel area
when a node is selected

6.2.1 Tutorial as pre-teaching and interpretive contract (Il Corollario)

- **Principle/requirement:** DP1 and DP2 imply that comprehension at scale requires staged entry and a legible grammar; DP9 requires that any editorial cue (e.g., size, highlighting, “suggested views”) is framed to avoid being mistaken for neutral truth.

- **Design choice:** treat onboarding as pre-teaching: users should learn the reading conventions before the first interpretive act, reducing cognitive overload and preventing early misreadings from becoming anchoring biases.

- **UI mechanism:** Il Corollario provides a tutorial that introduces (i) the debate constructs (subject/positions/arguments), (ii) AI-mediated layers (clusters and keywords), and (iii) interpretive cues (node size, hull regions, dashed links), explicitly stating how each should be read and what it does not mean.

- **Expected user value:** users can enter the KG with an operational mental model (“what counts as evidence, what counts as annotation, what is system-generated”), enabling more reliable sensemaking while keeping the interface accessible to non-expert audiences.

- **Limitations/trade-offs:** pre-teaching reduces misinterpretation but introduces front-loaded time/attention cost; in practice, users may skip onboarding. This makes in-context reminders (legend, tooltips, reversible views) necessary complements rather than optional extras.

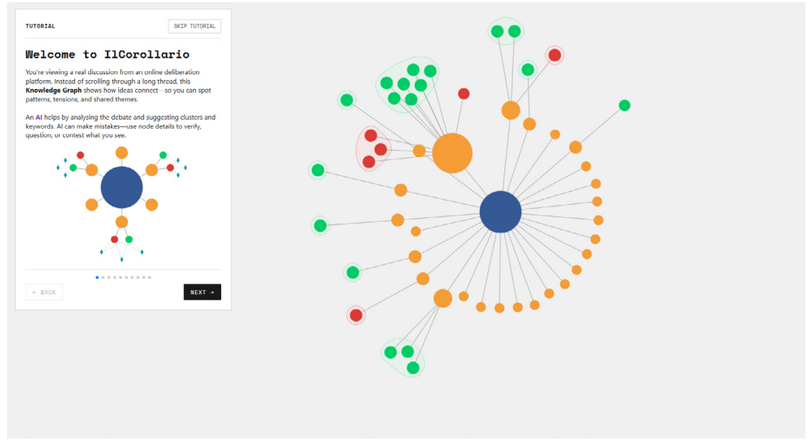


Fig.16 Il Corollario, tutorial first step, introducing the user to the platform

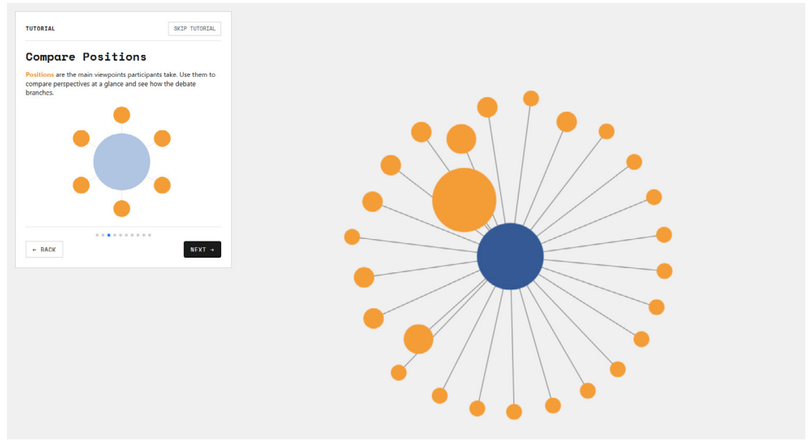


Fig.17 Il Corollario, tutorial third step, introducing the user to positions nodes

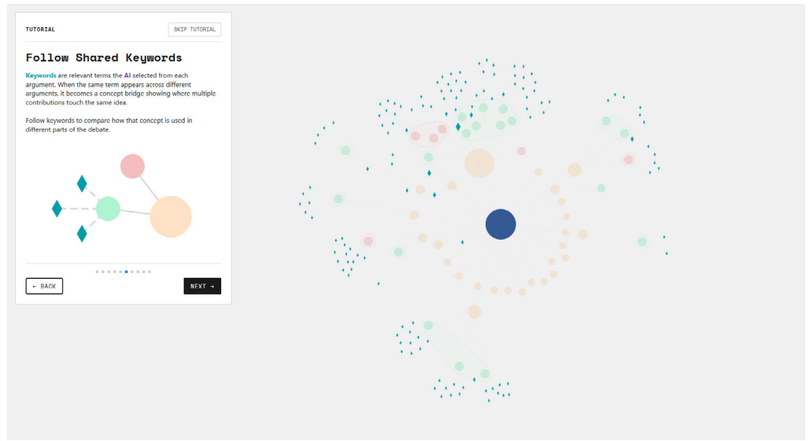
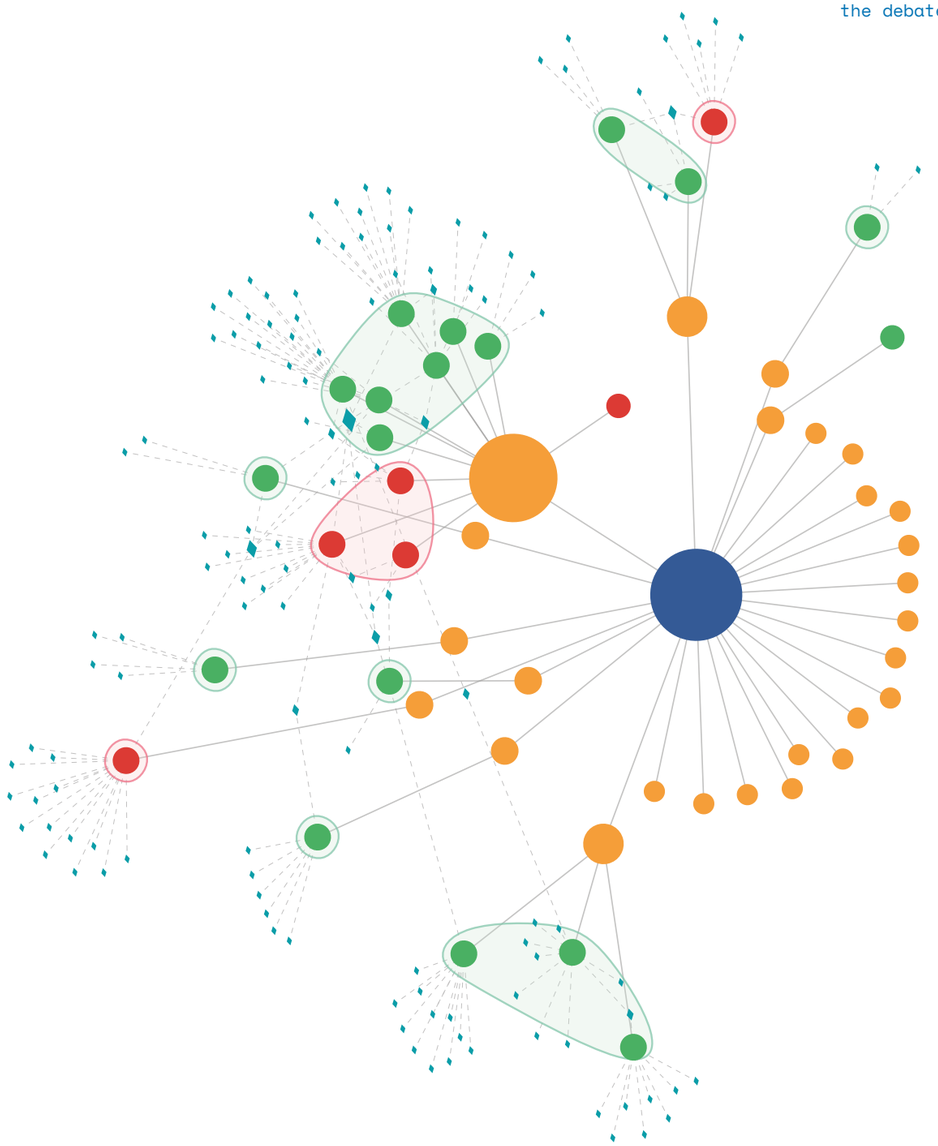


Fig.18 Il Corollario, tutorial sixth step, introducing the user to AI generated keyword nodes

6.2.2 Design rationale: “structural hierarchy as a representation of meaning”

- Principle/requirement: DP2 (legibility) + UIE.4 (user-friendly network exploration).
- Design choice: Replace visually uniform graphs with an explicit grammar that encodes debate roles, AI mediation, and analytic hierarchy preattentively.
- UI mechanism: Il Corollario adopts role-based colour, shape differentiation, degree-based sizing, stance-aware cluster areas, and link down-weighting for keyword relations.
- Expected value: Users can maintain global context while navigating a dense debate, because the graph communicates its semantics directly rather than delegating meaning only to click-through panels. This explicitly addresses the baseline issue where interpretation became “click-heavy” and cognitively disruptive.

Fig.19 Il Corollario, netowrk view showing the debate structure



6.2.3 Role legibility through preattentive encodings

- Principle/requirement: DP2 (legible visual languages) + DP3 (evidence binding) + DAV.6 (user-friendly viewpoints exploration).
- Design choice: encode deliberative roles as first-class visual categories so users can read the argumentative structure “at a glance.”
- UI mechanism: Node colour maps to debate roles (SUBJECT, POSITION, INFAVOR, AGAINST, KEYWORD/ENTITY), enabling recognition of the argumentative backbone. Node shape separates AI-extracted entities/keywords from discussion elements (diamonds vs circles), making the sociotechnical origin of information visible rather than implicit.
- Expected value: Users can distinguish participant-authored discourse units from AI-produced annotations, reducing the risk that extracted concepts are mistaken for primary deliberative claims. This aligns with the framework’s emphasis on treating mediation seams as communicable interface content rather than hidden implementation detail (DP4).
- Trade-off/limitation: Reliance on colour can disadvantage colour-blind users and low-contrast display contexts. The shape encoding mitigates this partially (ENTITY vs non-ENTITY), but stance distinctions (INFAVOR vs AGAINST) remain colour-dependent unless complemented by secondary cues (e.g., outline pattern or iconography). This should be acknowledged as an accessibility debt for future iterations.

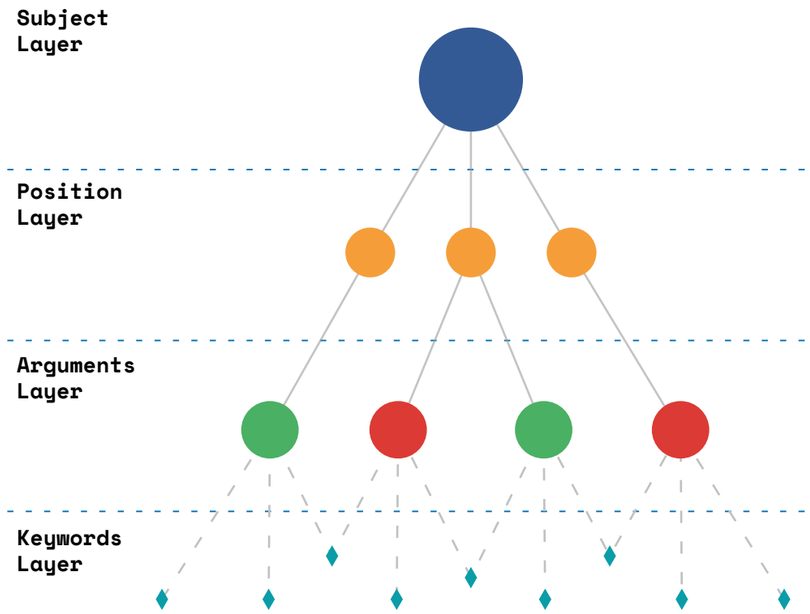
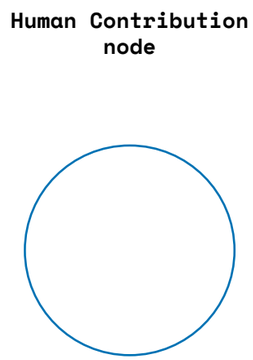
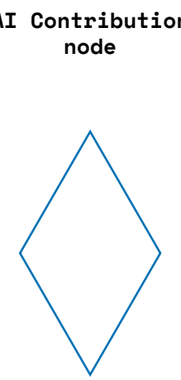


Fig.20 The components of the debated encoded in the platform





6.2.4 Visual hierarchy without conceptual dominance

- Principle/requirement: DP2 (legibility under cognitive constraints) + DP9 (rhetorical transparency of salience cues).
- Design choice: Use size to communicate connectivity while preventing a common KG failure mode: keywords/entities inflate degree and visually dominate the debate.
- UI mechanism: Il Corollario computes degree in two modes: Structural degree: connections among non-ENTITY nodes (debate structure). Conceptual degree: connections involving ENTITY/ keywords. Node radius uses separate scales so keywords remain small and discrete, while SUBJECT/POSITION retain a stable hierarchy.
- Expected value: node size becomes a guide to structural relevance (where the debate branches, where positions attract many arguments) without collapsing the interface into an “entity cloud.” This improves interpretability and keeps plurality legible (DP5) because minority positions are less likely to be visually erased by high-degree semantic hubs.
- Trade-off/limitation: Any salience encoding risks being misread as “importance” or “truth.” Il Corollario mitigates this through interpretive scaffolding in the tutorial (explicitly framing size as discussion connectivity rather than epistemic authority), but the risk cannot be eliminated—only governed through framing and reversible exploration paths.

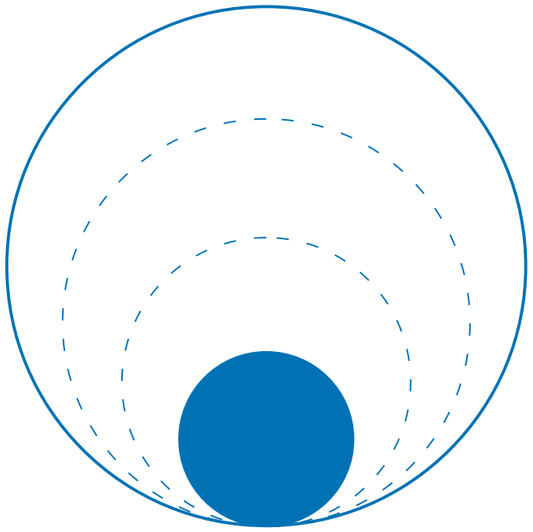


Fig.21 Node size of positions and keywords nodes in Il Corollario depends on n° of connections

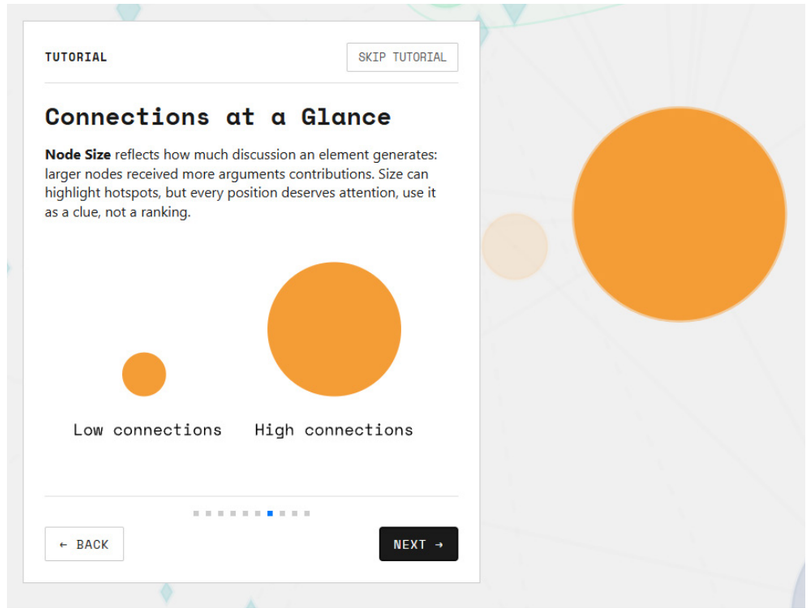


Fig.22 Il Corollario tutorial seventh step, explaining the difference between node sizes

6.2.5 Meaningful grouping: clusters as selectable “areas”

- Principle/requirement: DAV.1 (cluster viewpoints/arguments) + MA.4 (controversy/contestation cues) + DP5 (plurality and conflict).
- Design choice: Treat AI clusters not as hidden metadata but as a spatial layer that users can read and interrogate.
- UI mechanism: AI clusters are rendered as convex hull areas around member nodes; hulls are clickable and highlight members, turning themes into navigable regions. Cluster colouring is stance-aware: the tint reflects the balance of INFAVOR vs AGAINST inside the cluster (neutral if balanced; tinted if imbalanced).
- Expected value: Clusters become interpretable hypotheses about thematic organisation: users can see where argument communities form, and can check whether a theme is internally contested or aligned. This supports plurality-preserving sensemaking because disagreement remains visible within thematic groupings rather than being smoothed into a single narrative.
- Trade-off/limitation: Stance-aware tint can be read normatively (e.g., “red means bad”); it therefore needs careful legend framing and, ideally, a toggle to disable stance colouring for neutrality-sensitive settings. In the current prototype, mitigation is primarily communicative (legend + tutorial), which should be acknowledged as incomplete.

Due to the deliberative platform arguments contribution structure, users have to decide in which thread (IN FAVOUR/AGAINST) to place the contribution, and the thematic clustering done by the AI system cannot group arguments that are from opposing threads. When designing the platform I decided to imagine a future implementation where the grouping is made regardless of the user thread choice, in order to show how different contributions can talk about the same topic.

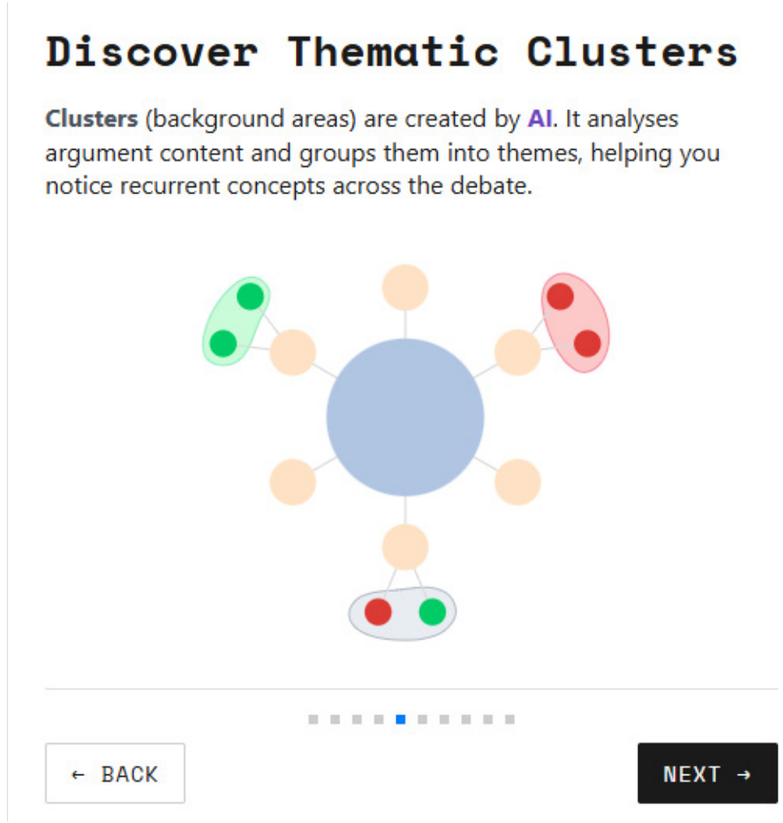


Fig.23 Il Corollario, tutorial fifth step, explaining what clusters are to the user and how are they generated

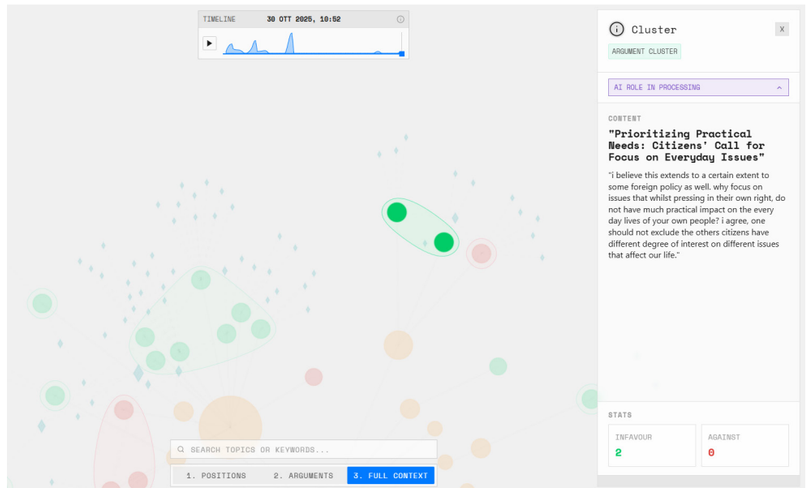
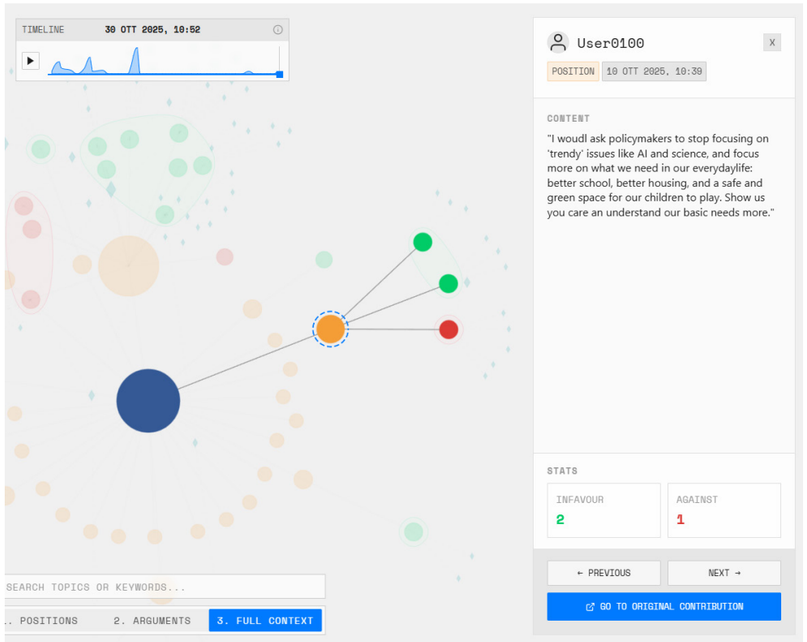


Fig.24 Cluster selected and highlighted inside Il Corollario, and the detail panel showing the cluster informations.

6.2.6 Noise control and relation legibility: differentiating “argument backbone” from “semantic annotation”

- Principle/requirement: DP2 (avoid overload) + DP3 (evidence binding) + UIE.4 (network usability).
- Design choice: establish a visual hierarchy between relations that define argumentative structure and relations that provide semantic context.
- UI mechanism: links involving keywords/entities are down-weighted (lower opacity + dashed strokes), explicitly separating conceptual annotation from the debate backbone and reducing hairball effects.
- Expected value: users can first read the deliberation as structure (subject → positions → pro/against arguments), then selectively inspect semantic bridges without losing the overall map. This supports the “structured legibility under overload” orientation developed in Chapters 4-5.

Fig.25 Position node selected and the detail panel with its informations



6.2.7 Interaction design as reading support (micro-interactions tied to encoding)

Several interactions in Il Corollario are tied to encoding because they function as “reading operators” over the visual grammar:

- Hover/selection highlighting reinforces the role encodings by isolating local neighborhoods (supports DP3 evidence-oriented inspection).
- Clickable cluster areas turn thematic grouping into an inspectable layer rather than a static label.
- Legend + tutorial scaffolding act as interpretive contracts, ensuring that visual cues (especially size and colors) are not mistaken for neutral truth—explicitly aligning with DP9.

Force-directed layouts can change node positions between sessions or after interactions, which can reduce spatial memory and reproducibility in reporting contexts. This is partially mitigated through export utilities and “fit/reset” navigation patterns (covered in 6.3/6.4), but remains a structural limitation of dynamic network layouts.

6.3 Supporting Collective Sensemaking through Network Visualization

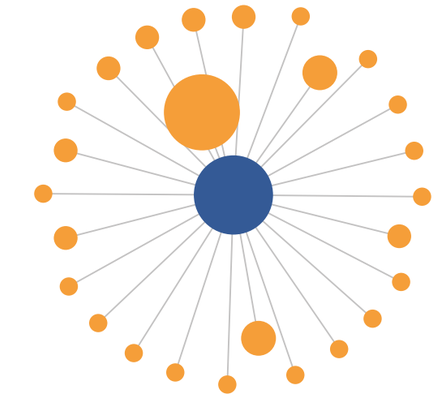
ORBIS positions the KG interface as a deliberative artefact that makes large-scale discussion inspectable: it should help users identify contention points, bridging concepts, peripheral voices, and temporal dynamics, enabling “discussion about the discussion” rather than isolated reading of statements.

Il Corollario operationalises this sensemaking role through an interaction grammar that combines progressive disclosure (to manage scale), structured interrogation pathways (to turn open exploration into repeatable analytic moves), and evidence-oriented inspection (to keep interpretations tethered to sources and to the system’s mediation choices).

6.3.1 Progressive disclosure as a civic navigation strategy

- Principle/requirement: DP1 (progressive disclosure under literacy constraints) and DP2 (legibility) emphasise that comprehension at scale requires staged, revisitable scaffolds rather than a single “full complexity” view.
- Design choice: treat complexity as something the user enters gradually, preserving orientation and reducing the cognitive penalty of dense networks.
- UI mechanism: Il Corollario implements depth navigation with three analytic depths—(1) Main positions, (2) Arguments, (3) Full context—plus manual drill-down that allows temporary expansion (e.g., reveal arguments from a selected position, then reveal entities from a selected argument).
- Expected user value: non-expert users can begin from a stable argumentative skeleton (topic → positions) while analysts can progressively activate richer layers (arguments and semantic bridges) without losing the macro-structure. It also gives the possibility to observe and analyse the discussion focusing on one position at a time or combining multiple positions without the visual clutter of the full context view.

Lvl. 1
Positions



Lvl. 2
Arguments

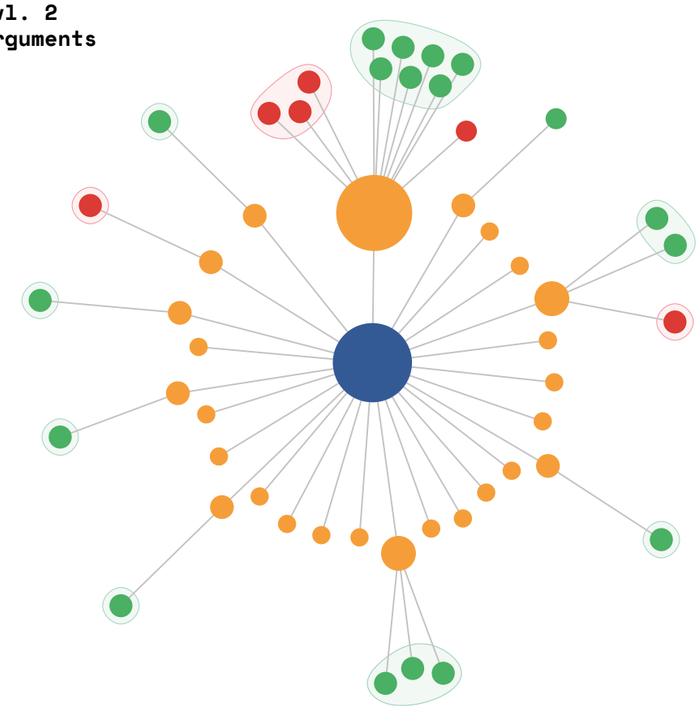


Fig.26 Two versions of the same debate filtered at different depth (1-2)

6.3.2 Structured interrogation pathways: from “open exploration” to repeatable analytic questions

- Principle/requirement: DP6 frames interaction as a repertoire of standardized interrogation pathways that make contestation and diagnosis actionable rather than purely interpretive. ORBIS also operationalises “points of interest” such as Most Contested positions as attention cues for deeper analysis.

- Design choice: embed question-shaped entry points into the interface so users can approach the same debate from different civic-relevant tasks (conflict, bridging, marginality), while preserving reversibility and inspection.

- UI mechanism: Suggested views provide task-specific lenses (e.g., Most Contested, Lone Voices) that reconfigure attention without rewriting the underlying graph. Search + minimap + zoom/fit utilities support: quickly locating a concept/position and re-establishing orientation after local inspection.

- Expected user value: Il Corollario makes sensemaking procedural: the interface offers interpretable paths that support both expert investigation and public-facing explanation (e.g., “show me where disagreement concentrates” or “who is isolated”).

- Trade-off/limitation: Suggested views introduce editorial force: they can be misread as “the system’s conclusion” rather than a lens. The mitigation is communicative and procedural—views must be clearly framed as starting points and remain reversible (return to neutral view; inspect the computation basis in node/edge neighborhoods). This aligns with DP9’s emphasis on rhetorical transparency for attention cues.

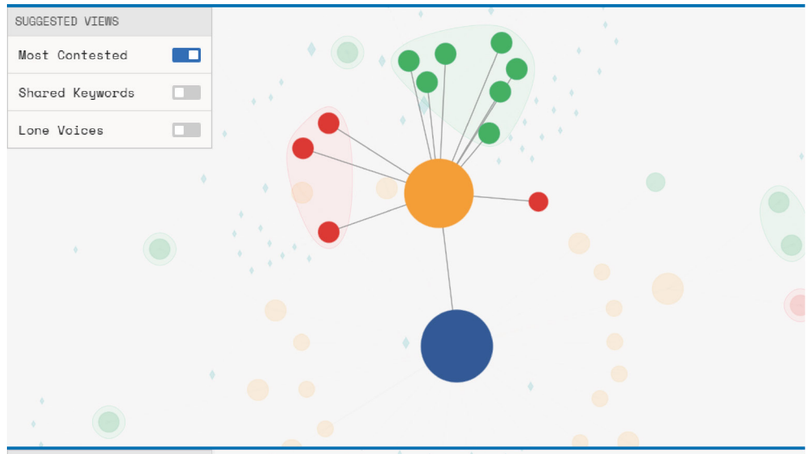


Fig.27 Suggested view section with the “Most contested” filter toggled

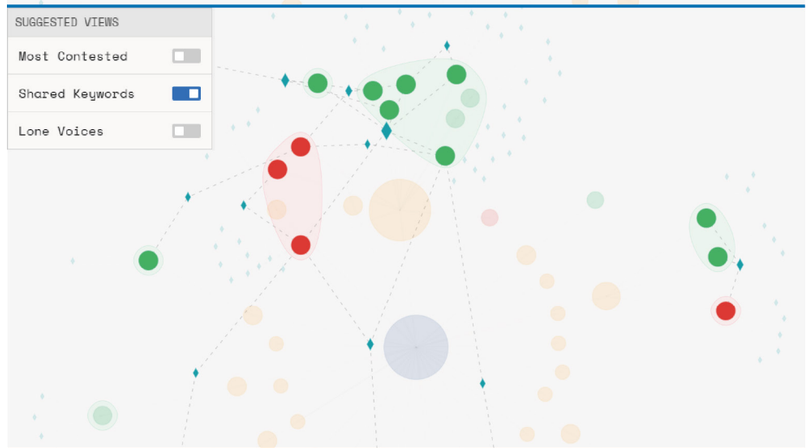


Fig.28 Suggested view section with the “Shared keywords” filter toggled

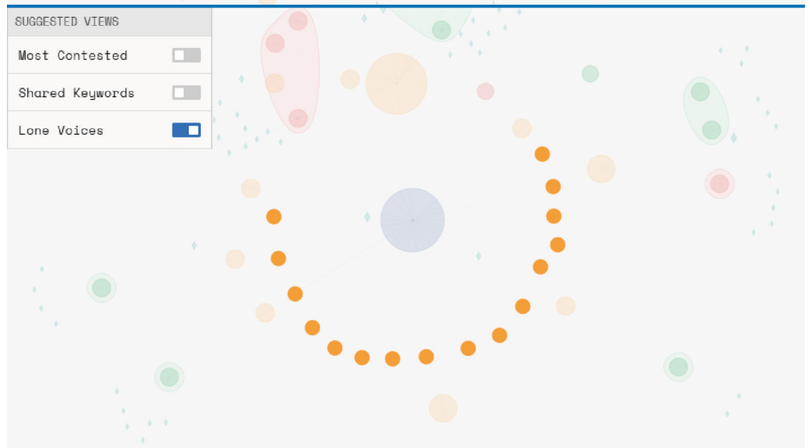
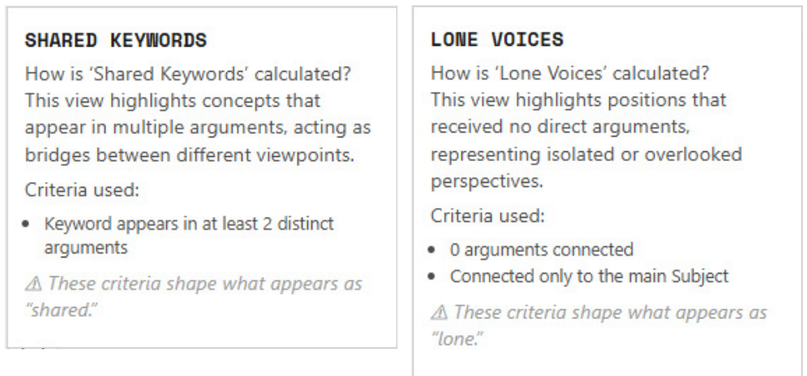


Fig.29 Suggested view section with the “Lone voices” filter toggled

6.3.3 Exposing lens criteria through hover tooltips

- Principle/requirement: DP8 and DP9 require that attention cues and analytic lenses are transparent: users should be able to see why the interface is emphasizing specific nodes.
- Design choice: expose “design knobs” not as configuration controls, but as auditable explanations of the criteria behind each suggested view—turning a black-box highlight into an inspectable lens.
- UI mechanism: when hovering suggested-view labels, Il Corollario shows tooltips that describe the selection logic (e.g., what “most contested” means operationally; what qualifies as a “lone voice”).
- Expected user value: users can interpret highlights as criteria-based suggestions rather than normative judgements, and can decide whether the lens is appropriate for their purpose.
- Limitations/trade-offs: tooltips increase transparency but do not guarantee comprehension; criteria should be expressed in plain language and, when possible, linked to inspectable evidence.

Fig.30 Tooltips that appear when the cursor is in hover on the suggested view’s filters



6.3.4 Evidence-oriented inspection: making verification the default reading practice

- Principle/requirement: The framework argues that explainability in civic contexts must support verification and challenge rather than only understanding; contestability needs concrete interaction mechanisms. ORBIS also requires reporting/summary outputs that maintain source links (RS.5) and treats KG as a tool for inspection and verification.
- Design choice: bind every interpretation to an inspectable trail: what you see in the graph should be traceable to which contributions and which relations support it.
- UI mechanism: Node selection triggers neighbor highlighting, turning “why is this connected?” into a perceptible relation set rather than an abstract claim. The details panel exposes node payloads (argument text, cluster summary, keyword value) and—where available—origin identifiers/links back to the source platform objects, aligning with RS.5’s requirement for source-link preservation.
- Expected user value: verification becomes low-friction: users can inspect evidence locally (neighbors and content) and, when needed, jump to the source context. This supports calibrated interpretation and reduces the risk of treating AI-derived groupings (clusters, keywords) as authoritative statements rather than analytic hypotheses.

Fig.31 Go to original contribution button to bring the user to the deliberation platform

6.3.5 Plurality-preserving sensemaking: bridging concepts without collapsing conflict

- Principle/requirement: DP5 treats visualisation as a civic instrument that must preserve difference and conflict rather than forcing coherence; ORBIS highlights overlap/disjoint exploration (DAV.4) and controversy detection (MA.4) as relevant analytic goals.
- Design choice: design for plurality as structure: enable users to see both (i) where positions are internally reinforced, and (ii) where cross-cutting concepts connect otherwise separate clusters of discourse.
- UI mechanism: Semantic bridges via keywords: users can highlight an ENTITY node and observe which arguments it links, operationalising “bridging concepts” as traversable connections rather than narrative claims.
- Expected user value: Il Corollario supports detecting connective tissue (shared concepts that could enable reframing or mutual understanding), without erasing the underlying polarity of support/attack relations.
- Trade-off/limitation: The keyword extraction system is AI based and multiple argument nodes could share a common keyword, but if the keyword represents a general concept the concept bridge created by the keyword links can wrongly induce the reader to think that arguments are semantically related where they just share a word.

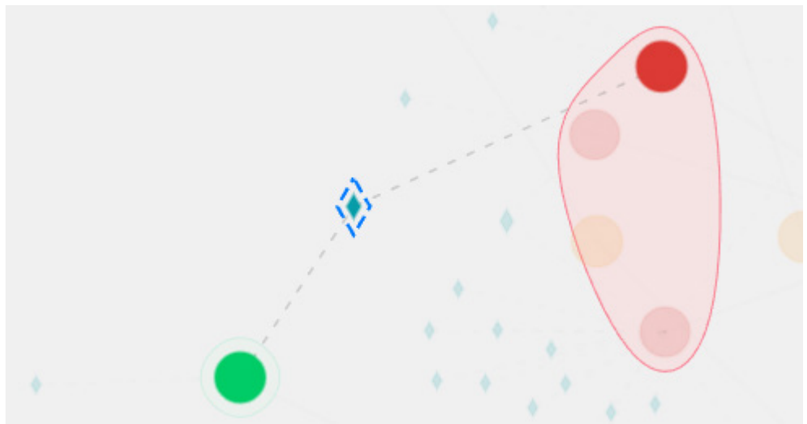


Fig.32 Keyword selected that connects two different arguments

6.3.6 Temporal sensemaking: deliberation as a process

- Principle/requirement: DP7 extends agency across time and phases; ORBIS explicitly targets mechanisms to capture deliberation evolution and dynamics and includes temporal exploration in the redesigned KG.
- Design choice: allow users to reconstruct how the debate developed (emergence of clusters, intensification of contestation) rather than only what the final topology looks like.
- UI mechanism: A timeline slider + playback filters the visible graph by timestamp to observe evolution. An activity “mountain chart” summarises participation density across time bins, providing a temporal overview that can be correlated with shifts in graph structure.
- Expected user value: moderators and analysts can triangulate structural patterns with temporal dynamics (e.g., spikes of activity preceding increased contestation), supporting process-oriented interpretation and report narratives that acknowledge deliberation as unfolding.
- Trade-off/limitation: In the current Il Corollario prototype, temporal analysis depends on availability and quality of timestamps in the ingested data. Where timestamps are missing or inconsistent, the timeline becomes a scaffold rather than a reliable analytic instrument.

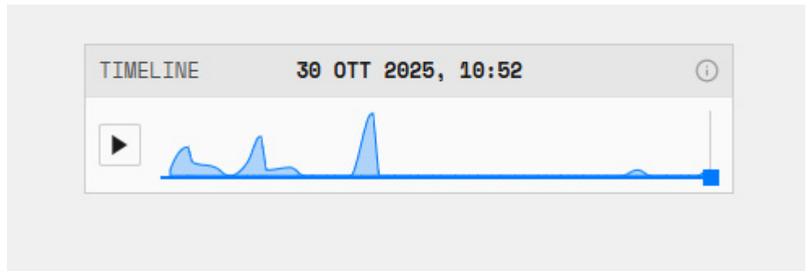


Fig.33 Timeline panel

6.3.7 Personal annotation as sensemaking infrastructure

- Principle/requirement: UIE.3 requires a personal space / idea resource pad; the framework also frames interaction as a means to externalise, revisit, and contest interpretations.

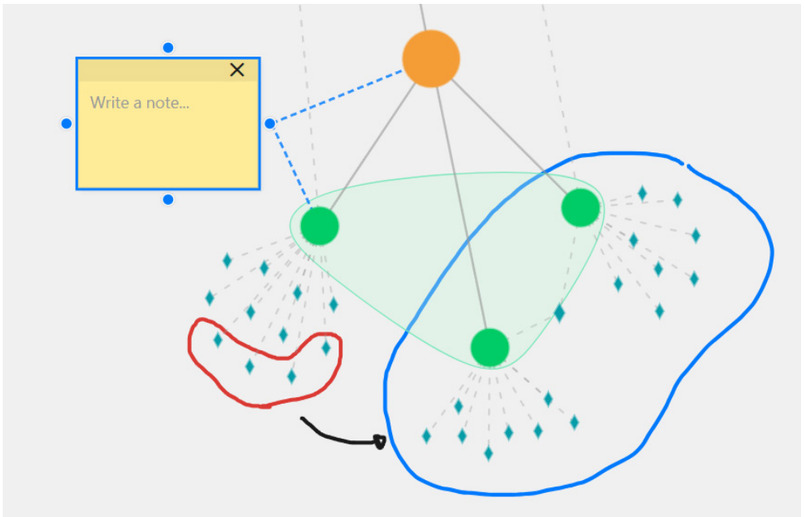
- Design choice: treat sensemaking as constructive work: users should be able to keep hypotheses, excerpts, and interpretive notes alongside the KG rather than outside it.

- UI mechanism: Il Corollario provides a post-it/notes space (plus lightweight markup tools) that users can place near relevant nodes/clusters to build a trace of interpretation that can later be exported into reporting workflows.

- Expected user value: this turns the KG from a passive visualisation into a working surface: users can develop explanations they can share (e.g., via export), while keeping a local record of why a certain pattern was interpreted as contested, marginal, or bridging.

- Trade-off/limitation: The “personal space” is primarily individual: unless synchronised or shareable, it supports personal analysis more than collaborative co-annotation. In future implementations it could become a collective analysis tool.

Fig.34 Personal notes’ postit linked to two different nodes and drawing function used to highlight an insight



6.3.8 Contestability loops: challenging AI-mediated layers

- Principle/requirement: In HCXAI terms, explainability in civic settings must support challenge as well as understanding; DP4 and DP6 imply that AI mediation should be visible and open to interrogation, not treated as an unquestionable analytic layer.

- Design move: implement contestability at the point of interpretation: users should be able to dispute AI-derived features where they encounter them (cluster membership, keyword selection), keeping disagreement as a legitimate interaction rather than an external support request.

- UI mechanism: Il Corollario exposes contest actions on AI-mediated objects: For keywords/entities, users can contest whether a concept is relevant or correctly extracted; for clusters, users can contest whether a contribution is appropriately grouped under a theme. These actions are presented as part of node/cluster inspection, maintaining the link between what is being challenged and the evidence context in which the judgement was formed.

- Expected user value: contestability supports epistemic agency: users can register disagreement with the system’s mediation, reducing the risk that AI groupings are interpreted as “the structure of the debate,” and enabling a feedback trace that can inform iterative refinement.

- Limitations/trade-offs: if contestation does not propagate back into the pipeline (e.g., re-clustering, updated keyword extraction), it functions as a communicative and governance signal rather than a corrective mechanism. In that case, the interface should frame contestability as “recording a challenge” rather than “fixing the model.”

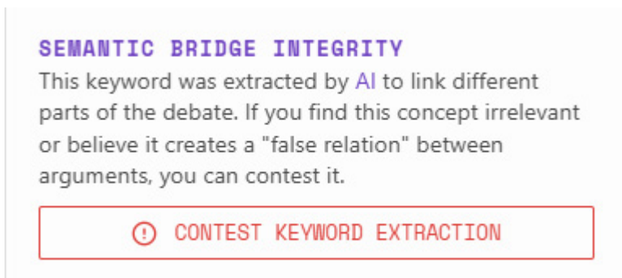


Fig.35 Ai-contribution contesting feature for keywords (also available for clusters’ nodes)

6.3.9 From data to knowledge: disclosure as a reportable artifact

- Principle/requirement: DP4 (AI role and limits) and DP8 (transformation transparency) require that users can understand what the system did to the data and what it did not do.
- Design move: introduce a dedicated disclosure panel that makes provenance and mediation explicit: where data came from, what transformations were applied, and what interpretive tasks remain outside the system's scope.
- UI mechanism: the "From data to knowledge" panel communicates: (i) data provenance (source platform/topic scope), (ii) two AI intervention layers (thematic clustering; keyword extraction/ bridging), (iii)non-capabilities / boundaries (what the interface does not claim to infer).
- Expected user value: users can calibrate trust in AI generated/ extracted content, avoiding over-reliance.
- Limitations/trade-offs: disclosure panels are only effective if they are discoverable and written for non-expert readers; if overly technical, they risk functioning as compliance text rather than actionable transparency.

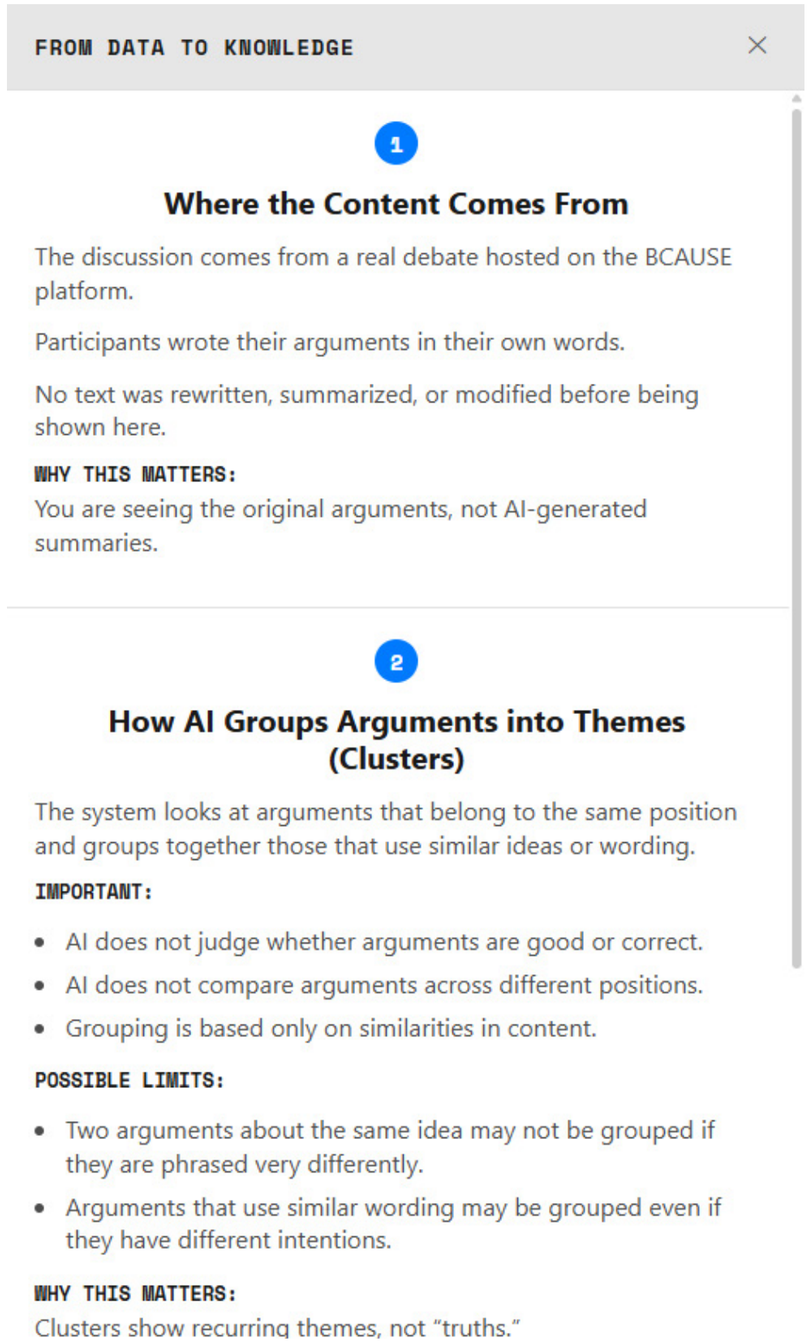


Fig.36 From data to knowledge panel showing the steps of data transformation from human contribution to the AI pipeline

6.4 Integration into the ORBIS Reporting Ecosystem

ORBIS operationalises visual analytics through a flexible reporting system: moderators can generate time-snapshot reports at different moments in a discussion, producing a historical archive of report states that can be revisited and compared. Reports are delivered as an interactive dashboard composed of modular widgets, accessible to participants, and exportable (single widgets as PNG; full dashboard as PDF), supporting documentation and reuse in facilitation and reporting practices. In this context, Il Corollario was conceived as an additional interpretive lens for navigating deliberations as networks—complementary to the existing "data-rich" analytics and more visually oriented exploration interfaces described in ORBIS.

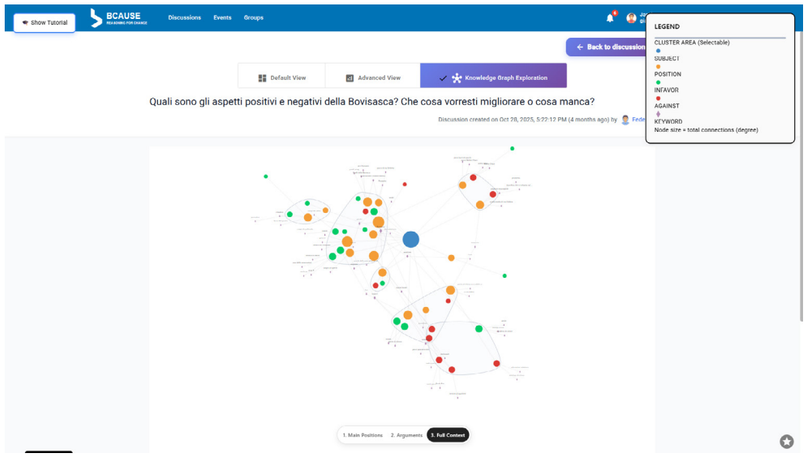


Fig.37 Early implamentation of the KG inside the BCause reporting system

6.4.1 Implementation status and placement in BCause reports

Due to implementation timing constraints relative to the ORBIS project schedule, only an older, feature-reduced version of Il Corollario was integrated into the BCause reporting system. Development continued beyond that initial deployment, and the newer interaction and transparency features described in this chapter remain technically implementable in future iterations.
In the BCause reporting interface, the KG view was positioned as a third tab alongside the existing report views (Default and Advanced), providing a dedicated space for network-based exploration without replacing the

standard report reading workflow. This placement reflects a design intent aligned with the framework: keep the KG as an optional but accessible sensemaking infrastructure rather than a mandatory gateway for understanding the report.

Preliminary platform-level feedback (not KG-specific)

The reporting system included a lightweight rating module where users evaluated the report on a 1–5 scale. The available averages are reported below.

Tab. 10 Evalutaion module inside BCause reporting platform

Item	Metric (1–5)	Average
1	The clarity and quality of this output	3,47
2	Easiness of use	3,53
3	Support for discussion (present and future)	3,8
4	Intention to use it again	3,97

These values should be interpreted as partial, platform-level signals rather than an evaluation of Il Corollario: the KG tab was introduced at the end of the implementation timeline, therefore it likely had limited opportunity to influence users' overall assessment of the reporting experience. In methodological terms, they provide context for the maturity and perceived usability of the reporting container in which Il Corollario was embedded, but they do not constitute evidence about the KG's communicative effectiveness or democratic adequacy.

6.4.2 Bridge to Chapter 7: validation as the dedicated evaluation phase

For these reasons, the evaluation of Il Corollario is treated explicitly in Chapter 7 (Validation. User and Expert Interviews), structured to separate (i) the user experience and usability from (ii) a final expert validation against the full set of requirements and design principles. Chapter 7 will cover: objectives and rationale; participant selection and expert profiles; user interview protocol (tasks + evaluation); analysis; expert interview protocol against requirements; analysis and thematic coding; and key findings with implications for design refinement.

7.0
VALIDATION. USER AND EXPERT INTERVIEWS

7.1 Objectives and Rationale of the Evaluation

This chapter validates Il Corollario as a communication design artifact that supports the interpretation of an AI-mediated deliberation. Coherently with the framework introduced in Chapters 5, the evaluation does not ask whether the AI is "right" in an absolute sense, but whether people can understand and scrutinize the mediation pipeline that transforms many contributions into collective representations (e.g., positions, argument structure, clusters, and keywords), and whether the interface makes this transformation legible, accountable, and contestable.

The evaluation therefore focuses on four framework-level dimensions that Il Corollario operationalizes through narrative framing, visual encodings, and interaction design: transformation transparency, plurality-preserving sensemaking, agency/contestability mechanisms, and evaluation beyond trust (i.e., supporting verification behaviors rather than eliciting passive acceptance). In practical terms, the goal is to verify that users can move from overview to evidence, interpret AI-generated elements as interpretive constructs (not neutral facts), and maintain the ability to compare perspectives and challenge questionable outputs.

A further rationale concerns real-world constraints, validation is thus also used to identify which requirements were fully met, which were approximated/adapted, and which remained out of scope for this iteration, so that design implications and next steps are grounded in observed usage and expert critique.

7.1.1 Two evaluation moments: user tests + expert validation against requirements

Validation was organized into two complementary moments:

- User tests (task-based interviews) assessed learnability, comprehension, and interaction costs in realistic exploration scenarios. This moment prioritizes practical sensemaking: whether non-experts can navigate the graph, interpret node/link meanings, connect AI constructs back to underlying arguments, and recognize how design choices (filters, node size, suggested views) shape reading.

- Expert validation (requirements-based survey/interviews) assessed Il Corollario against the requirements and design principles derived from the literature and the ORBIS framing. Experts were asked to perform targeted checks (tutorial clarity, provenance distinction, parameter inspectability, uncertainty disclosure, cross-checking via alternative views, traceability, contestability) and then rate each criterion on a 1–10 likert scale, complemented by open-ended feedback. This second moment functions as a structured, auditable evaluation of "democratic adequacy" at the level of the framework.

7.2 Participant Selection and Expert Profiles

A total of 10 adult participants (P01–P10) took part in the user study. The sample was intentionally oriented toward the tool's intended audience—non-expert readers of civic debates—while including one designer (P07) to capture a more "expert eye" on interface reasoning and communication choices.

Tab. 11 Age and number of participants to the user test

Age range distribution	number of participants
18–25	3
26-35	5
56-65	2

Before the User to each participant was asked the familiarity level both with data visualization and Artificial Intelligence, the engagement with online discussions/participatory platforms and the interest in civic/social/political topics.

- Familiarity with data visualizations: 6 moderately familiar, 3 slightly familiar, 1 not familiar
- Familiarity with AI: 7 slightly familiar, 3 moderately familiar
- Engagement with online discussions/participatory platforms: 6 rarely, 4 never

- Interest in civic/social/political topics: 6 very interested, 4 moderately interested

Expert participants (N = 10)

12 experts completed the requirements-based validation. The recruitment targeted complementary perspectives relevant to the framework: communication/design research, data visualization, AI transparency/XAI, and (where possible) deliberation/civic tech knowledge. Expert profiles were collected as short self-descriptions (with no personal identifying data stored), and results are reported only in aggregated form.

To characterize the panel's methodological readiness for the task, experts self-rated familiarity on 1–10 scales:

- Familiarity with graph-based visualizations: mean 9.0/10 (high)
- Familiarity with XAI / AI transparency tools: mean 7.5/10 (moderate–high)

This composition supports the purpose of the expert phase: not "general usability" judgment, but a requirements-level assessment of whether Il Corollario communicates provenance, uncertainty, parameters, plurality, traceability, and contestability in ways consistent with the design framework.

7.3 User Interviews

The user evaluation combined semi-structured interviews with a task-based usability/interpretation test to observe how non-expert readers make sense of a deliberation through Il Corollario. The focus was not only on interaction fluency, but on whether users could (i) build a correct mental model of what the visualization represents, (ii) move from overview → evidence (graph → original contributions), and (iii) recognize and critically interpret AI-mediated constructs (clusters/keywords) as non-neutral interpretive layers.

7.3.1 User Interview Protocol with tasks and evaluation

Each session followed a consistent structure:

1) Consent + short background questionnaire (age range, familiarity with data visualizations and AI, civic interest, participation in online discussions).
2) Guided onboarding through the tutorial, followed by 4–5 minutes of free exploration (explicitly requested before the first task).
3) Task sequence (T01–T13), moderated with a light think-aloud approach and short probing questions ("What are you looking at?", "What makes you think that?", "Where would you verify this?").
4) Post-test questionnaire: System Usability Scale (SUS) + optional open comment.

The 13 tasks were designed to cover the core interpretive moves required by the framework (navigation, evidence access, transparency of mediation, plurality cues, and contestability). For readability, tasks can be grouped into six functional clusters.

Tab. 12 User test's task clusters and evaluation goals

Task cluster	Tasks	What it evaluates
Mental model & orientation	T01–T02	Ability to explain the representation and identify discussion topic/context
Core graph reading	T03–T04	Understanding of positions and clusters; ability to navigate to relevant areas
AI constructs comprehension	T05	Meaning of keywords/bridges; ability to verify them in source text
Traceability & alternative readings	T06–T07	Timeline reading; switching perspectives through suggested views
Annotation as civic reading support	T08–T09	Personal notes + drawing as tools for reflection and recall
Cross-ecosystem verification	T10–T12	Bridging between BCause report views and the KG (how to re-find evidence)
Contestability & agency	T13	Recognizing AI-used elements and what to do when disagreeing

For each task four operational measures were collected in structured sheet:

- Completion: Yes / Partial / No
- Errors: No / Marginal / Yes
- Hints given: Yes / No (when the moderator provided directional support)
- Confidence rating: 1–5 (star scale), collected immediately after task completion
- Short qualitative notes were recorded continuously (breakdowns, misunderstandings, interface elements missed, and spontaneous design suggestions).

7.3.2 Data Analysis

Quantitative analysis

Two quantitative layers were used to characterize performance and identify friction points:

Task performance (130 observations = 10 participants × 13 tasks):

- Completed: 96.15% (125/130)
- Partial: 3.08% (4/130)
- Failed: 0.77% (1/130)
- Tasks with errors (Yes or Marginal): 16.15% (21/130)
- Tasks requiring hints: 15.38% (20/130)
- Mean confidence: 4.59/5 (SD 0.83)

SUS (N=10): mean 82.5/100 (SD 8.33, range 72.5–92.5), indicating a generally strong perceived usability, with variance that helps interpret individual friction points.

To keep interpretation grounded, the analysis highlighted tasks where performance systematically degraded (lower completion and/or higher

errors/hints), as these tasks tend to correspond to “meaning-critical” requirements (e.g., evidence bridging, interpretive seams, contestability).

The most demanding tasks were:

- T12 (bridge from BCause view → locate the same content in the graph): highest error rate (50%) and hint rate (40%).
- T05 (keywords/bridges + verification in text): one failure; elevated errors/hints (30% / 30%).
- T13 (AI elements + contestability action): lowest completion (80% fully completed, 20% partial), indicating that “agency against AI decisions” must be made more discoverable and self-explanatory.

Qualitative analysis and thematic coding

Qualitative data came from:

- the interview extraction document (key quotes/insights per participant + candidate codes),
- the structured task notes (observed breakdowns and spontaneous suggestions),
- optional SUS open comments.

Coding followed a compact thematic approach designed to avoid an excessive number of labels: Familiarization and initial coding: key insights were tagged using an initial pool of candidate codes derived from the extraction sheet, translation for reporting (English): since interviews and notes were collected in Italian, the final code labels and representative excerpts were translated into English, preserving intent rather than literal phrasing,triangulation: themes were cross-checked against quantitative signals (tasks with higher errors/hints and SUS variance) to avoid over-weighting isolated quotes.

The resulting codebook contains 7 themes, each directly actionable for design refinement:

- C1 Visual sensemaking efficiency (overview value vs linear reading)
- C2 Graph↔text evidence bridging (need for strong pathways back to original contributions)

- C3 AI seams & uncertainty comprehension (clusters/keywords as interpretive constructs)
- C4 Contestability as civic agency (how to challenge AI outputs and what that action means)
- C5 Plurality & minority visibility (reading "lone voices" and contested areas as democratic signals)
- C6 Rhetorical cues & bias prevention (node size/colors and perceived salience/neutrality)
- C7 Discoverability & accessibility friction (contrast, button salience, small text, missed affordances)

7.4 Expert Interviews

To complement the user tests (focused on usability and sensemaking), the second validation moment consisted of an expert evaluation against the requirement catalogue (Ch.5) and its operationalisation in Il Corollario (Ch.6). This step was designed to assess democratic adequacy at the level of the mediating layer—i.e., whether the interface makes AI-mediated deliberation legible, verifiable, plurality-preserving, and contestable, rather than merely "trustworthy."

7.4.1 Expert Interview Protocol against Requirements

The expert validation was conducted as a structured online questionnaire (asynchronous expert "interview"), combining (i) guided interaction tasks on the prototype and (ii) quantitative ratings plus open-ended prompts. This format enabled experts to evaluate the platform directly while performing the verification actions implied by the requirements (e.g., checking provenance, inspecting AI layers, cross-checking views), rather than responding abstractly. For each question a custom-made GIF showing the specific action was made to avoid going back and forth from the survey to the platform.

Structure of the survey: After a brief profile section (expertise + self-assessed familiarity with graph visualizations and XAI tools), experts were asked to:

1) Explore the interface freely to form an initial mental model (tutorial + legend + navigation scaffolds).
2) Execute a sequence of tasks, each followed by a 1–10 agreement rating tied to a requirement-oriented statement.
3) Answer open-ended questions on (a) the single feature that most improves transparency/verification and (b) the feature most at risk of misinterpretation or bias.

How tasks were mapped to requirements:
Each task was designed to operationalise one or more requirements (R01–R17) from Chapter 5, and the consolidated design principles (DP1–DP9) that frame the platform mechanisms in Chapter 6. For reporting and interpretation, items were grouped into the following requirement clusters:

- Entry scaffolding and complexity management (DP1–DP2; R01–R02–R08): tutorial, legend, and depth-based progressive disclosure to reduce overload while keeping expert pathways open.
- Legible visual grammar for deliberation structure (DP2; also supporting DP5): role-encoded node/edge grammar meant to make the debate backbone readable before click-through inspection.
- Evidence binding and provenance marking (DP3; R03–R04): verification flows that keep higher-level constructs anchored to arguments and distinguish human-authored vs AI-mediated layers (e.g., inspection, link hierarchy, provenance cues).
- Seamful AI mediation and transformation transparency (DP4 + DP8; R05 + R13): disclosure of what the system does (clustering/keyword extraction) and what it does not claim, through dedicated panels and framing of uncertainty/limits.
- Structured interrogation and rhetorical accountability (DP6 + DP9; R07 + R16): filters and alternative "views" intended to support cross-checking and reduce single-framing effects; explicit framing of salience cues (e.g., node size) as editorial/interpretive rather than "truth."
- Traceability, contestability, and governance signals (DP7; R09 + R17; partial R14–R15): temporal navigation (revisiting states), and

in-situ contest actions that allow users to challenge AI-derived layers at the moment of interpretation.

Requirements not fully covered by the expert survey: Consistent with the prototype's scope, some requirements from Chapter 5 were not directly testable (or only testable via weak proxies). For instance: upstream reflection scaffolds during contribution drafting (R06), and full workflows for negotiated "common ground" revisions (R12) are not core functions of the current KG interface; similarly, robust manipulation defenses (R11) exceed what can be validated in a single prototype view without broader platform governance features. These gaps are treated explicitly in Chapter 7 as scope-limits rather than evaluation failures.

7.4.2 Data Analysis and Thematic Coding

Ratings: Quantitative data were analysed using descriptive statistics (mean, Standard Deviation) for each survey item. Since most items were designed as 1:1 probes of requirements, results are reported both as per-item ratings and as requirement coverage through the mapping table.

Open-ended feedback: Given the brevity of responses, open answers were not subjected to formal thematic coding. Instead, they were synthesised discursively by identifying repeated points of agreement across experts.

Most cited "transparency/verification enablers" were onboarding and disclosure mechanisms: the tutorial, the legend/how-to-read, and the explanatory panels that clarify the AI role in processing and "From data to knowledge." Several experts also explicitly valued the Contest AI affordance as a verification/agency instrument (i.e., transparency is not only disclosure but also the ability to challenge).

Most cited "misinterpretation/bias risks" concentrated on AI-generated clusters and keywords (risk of over-reading them as objective structure), and on visual salience cues (node size and, to a lesser extent, colour/styling) that can steer attention or appear as normative judgment if not continuously framed. A smaller set of comments pointed to discoverability issues: when explanatory content exists, the risk is not its absence but whether it is found at the moment of interpretation.

Rating results: Across the 19 rated items, the mean evaluation was 8.36/10, indicating overall strong perceived democratic adequacy and usability for expert reading.

Highest-rated items (strengths):

- Tutorial onboarding (R01): 9.4 ± 0.8
- Legend / visual grammar comprehension (R08): 9.4 ± 1.1
- Recommendation as a complementary view in real deliberation: 9.0 ± 1.2

Lowest-rated items (priority improvement zones):

- Cross-checking via alternative views (R11): 7.5 ± 2.1
- Provenance clarity (human vs AI links) (R04): 7.7 ± 2.6
- Inspectability of parameters behind "contested / lone voices / shared keywords" (R15): 7.9 ± 2.6
- Suggested views as comparison tools (R07): 7.9 ± 1.8
- Timeline / traceability over time (R09): 8.1 ± 2.6

Notably, several of the lower-rated items also show high dispersion, suggesting that adequacy depends strongly on discoverability and on experts' interpretation of how "auditable" the mechanism feels (e.g., provenance cues and parameter inspectability). The high dispersion value could also be related to the effective time spent by the expert exploring the platform before answering the questionnaire (it was requested to explore for at least 2 min).

7.5 Key Findings and Implications for Design Refinement

7.5.1 Key findings from user interviews

Overall, the user study indicates that Il Corollario supports a fast, orientation-first reading of deliberations, with high perceived usability and high task success. Across 130 task observations (10 users × 13 tasks), users completed the vast majority of tasks successfully (96.15% completed, 3.08% partial, 0.77% failed) with high confidence (mean 4.59/5). Post-test usability perception was strong (mean SUS = 82.5/100, range 72.5-92.5). These outcomes suggest that the interface is generally legible and learnable even for participants who reported only slight-to-moderate familiarity with AI and data visualization.

However, the study also identifies three meaning-critical friction points where usability intersects with democratic adequacy: (i) evidence bridging from graph to source text, (ii) interpretation of AI-mediated layers (keywords/clusters) as "seamful" constructs, and (iii) discoverability of contestability actions as a form of civic agency. These challenges concentrate on a small subset of tasks (notably those that require verification or challenging the AI), and they align with the framework's emphasis on explainability as a verification and contestation-oriented practice, rather than generic "trust."

The graph enables "general ↔ particular" sensemaking, reducing the cost of orientation

Participants consistently described the network view as a cognitive shortcut compared to scrolling a linear discussion: the graph makes it easier to identify the structure of the debate (positions, where discussion concentrates, peripheral areas), and to decide what to read next. Users explicitly framed the visual layer as a way to avoid the burden of reading everything and "remembering everything," using the map as a selection mechanism before deep reading. This indicates that Il Corollario's overview and progressive disclosure logic effectively supports orientation at scale, even for participants with limited prior exposure to civic platforms.

At the same time, users highlighted that visual comprehension has a threshold: for small quantities, they "count" visually, but for larger graphs they rely on numeric summaries to avoid fatigue. This confirms the value

of combining spatial reading with summary statistics as complementary cognitive supports.

The interface already performs well as an orientation layer; design refinement should protect this strength by reducing avoidable friction (especially around returning to evidence and resetting context), so that "overview → inspection" remains fluid rather than becoming click-heavy. But further testing is needed to understand how the platform behaves with larger discussions.

Users treat the graph as a filter, but still need strong pathways back to evidence

Il corollario provides to the user the possibility to read the full text contribution of each node inside the side-panel, and if the reader wants to trace back the contribution to the original platform a button called "go to the original contribution" is available at the bottom of the text. But going from BCause to the Network visualization was not immediate, just few users used the searchbar to write a snippet of the text to be sure to find the correct node. This is visible in Task T12, where error and hint rates peaked

If we consider Il Corollario as a stand alone artifact, where there is no relation between the deliberation platform where users created their contribution, there are no problems in reading the content of each visualization element. But if we consider the platform as an assisting visualization that supports the deliberation platform, the link between the two needs some refinements to be more clear and accessible.

AI layers are generally understood, but keywords need contextualization and "seams" must be more actionable

Most participants correctly recognized that clusters and keyword nodes are AI-derived interpretive layers (and not human-authored statements), especially when tutorial framing was attended to. Users could also interpret the function of clusters as thematic grouping, and some participants showed an advanced understanding of confidence as a proxy for grouping reliability/homogeneity.

Where comprehension weakens is at the level of semantic verification: Task T05 (keywords/bridges) showed the only failure case and elevated errors/hints. Users wanted the interface to explicitly show why a keyword matters: where it appears in the text, in what sentence, and with what contextual

meaning. Without this, keyword nodes risk being read as opaque labels that are hard to judge, or as weak evidence that nonetheless shapes navigation.

For AI-mediated objects, Il Corollario should provide "micro-evidence" that supports immediate evaluation: keyword highlighting within text, short extraction context, and clearer scaffolding around what the AI did and did not do. This translates seamful design from a disclaimer into a practical verification aid.

"Most contested" and "lone voices" are read as democratic signals, but visual rhetoric can bias interpretation

Participants naturally interpreted "most contested" as a civic-relevant hotspot and "lone voices" as either (a) under-heard citizens or (b) minority/low-engagement perspectives, both readings treat the interface as an instrument for plurality-aware sensemaking. This supports the idea that network patterns can surface civic dynamics that linear interfaces hide.

At the same time, users also pointed out that visual cues (color and size) can act as rhetorical emphasis. One participant explicitly proposed a more "neutral" view (e.g., temporarily hiding pro/against coloration) to reduce early anchoring effects while reading. This indicates an important tension: salience cues are useful, but they must be framed and controllable to avoid being mistaken for neutral truth.

A possible future implementation could be to provide a neutral reading mode (or staged reveal of stance cues) so users can first inspect content without being visually steered, then re-enable stance overlays when comparing positions. This preserves plurality while mitigating premature framing.

Contestability is valued, but still not sufficiently discoverable as an interaction "grammar"

Participants expressed strong approval of the possibility to contest AI outputs, describing it as a mechanism that returns power to citizens and "rehumanizes" interpretation. Yet Task T13 had the lowest full completion (only 80% fully completed; 20% partial), suggesting that the action is not always discovered or understood in time, even when users agree with the principle.

This points to a design gap between having contestability and making it an obvious, situated practice: users should encounter contestation as part

of reading (when an AI object is inspected), with clear explanation of what contesting means (recording disagreement; potential feedback loop; not necessarily immediate correction).

A consequential design change would be to increase the visibility and clarity of contestability at the moment of interpretation (in the details panel for clusters/keywords), and frame it explicitly as a civic right: "Challenge this AI grouping/extraction."

7.5.2 Expert validation key Findings

After the Analysis and the examination of the expert validation results i created a list of key findings and a final table with all the questions mapped to the requirements set defined in chapter 5, with some consideration regarding which requirements were fully met, which requirement was adapted and need some changes and which requirement were excluded because not implemented in the platform but that still need attention for future implementations.

Key findings:

1) Entry scaffolding and legibility are strong (R01, R02, R08).
   The tutorial and legend/visual grammar are the highest-rated components (Q01 and Q03 both mean 9.3/10), supporting the framing that progressive disclosure and legible graph conventions are prerequisites for civic interpretability in complex deliberation maps.

2) Evidence-oriented inspection is credible but not uniformly experienced (R03, R04). Evidential anchoring (Q04, 8.2) is solid, consistent with the "details-on-demand" and inspection loop described in Chapter 6. However, provenance marking (Q05, 7.5, high variance) is one of the weaker points. This aligns with expert comments reporting residual ambiguity about what is "in the original data" versus "AI-generated" (especially around clusters) and with discoverability issues around interpretive disclosures.

3) "Seamful" AI communication works conceptually, but clarity depends on discoverability (R05, R13). Uncertainty/limitations communication (Q06, 8.1) and the "From data to knowledge" disclosure (Q14, 8.1) indicate that experts generally recognise the system's seamful framing (AI as fallible, interpretive mediation). This corresponds to the Chapter 6 design choice of making

AI boundaries and non-capabilities explicit. At the same time, qualitative feedback highlights a concrete UI risk: when the disclosure entry point is hard to find, the seamful framing becomes structurally correct but practically invisible. The “From data to knwoledge panel” needs to be more visible and explicit in order to be found and read at the right moment before exploration and not after.

4) Plurality representation is validated (R10), while anti-manipulation support is moderate (R11). Disagreement/minority visibility (Q10, 8.3) suggests that the interface’s core plural structure (positions + in-favour/against contributions + optional “lone voices”) supports the “plurality-preserving” requirement. The “cross-checking against framing/manipulation” item (Q11, 7.4) is the lowest mean score, suggesting that providing multiple views is helpful but not yet sufficient to consistently make rhetorical risk legible as an explicit user practice (as anticipated by Chapter 5’s warning about narrative/visual persuasion).

5) Editorial cues are mostly understood, but node salience remains a perceived bias risk (R16). The node-size explanation item (Q12, 8.3) indicates that, after reading the framing, experts generally interpret size as a discussion-connectivity cue rather than truth/value. This matches the explicit stance in the tutorial/visual rationale in Chapter 6. Yet multiple experts still cite node size as the single most bias-prone cue, reinforcing that transparency reduces, but does not eliminate, salience misreadings. A possible solution would be to reduce the size-connections ratio to make it less prominent.

6) Contestability is appreciated, but “revision workflow” is only partially instantiated (R17). Experts rate the contest action positively (Q15, 8.3), and the implementation indeed places contest actions in-situ within node inspection. However, the current prototype’s contestation is not linked to an actual revision pipeline or auditable history; it is implemented as a “feedback recorded” interaction without a traceable workflow (in fact that functionality is a prototype, not really implemented, it wasn’t disclosed in the survey question to avoid influencing the results). This is consistent with expert feedback asking for a way to support contestation with comments/annotations (i.e., structured rationale, not only a click).

7) Adoption potential is high (global items): Experts report high overall fluency (Q17, 8.3) and democratic adequacy (Q18, 8.2), with a strong recommendation score (Q19, 8.8), suggesting the interface is perceived as a valuable complement to linear discussion reading, coherently with the intended role of the KG as an optional sensemaking layer.

7.5.3 Design refinement implications (expert validation):

Below are the refinements that are directly supported by the expert feedback and the implementation constraints documented in Chapter 6.

Make interpretive disclosures harder to miss (R05, R13, R15, R16)

- Problem signal: outliers and qualitative feedback indicate that users may not discover the “From data to knowledge” panel and the suggested-view criteria tooltips.
- Refinement: elevate disclosure affordances from “available” to “ambient” (e.g., persistent label, clearer iconography, onboarding step that explicitly points to these controls, or inline disclosure microcopy near suggested views). This strengthens the practical fulfilment of seamful design and parameter legibility.

Clarify cluster provenance and reduce “AI-style” verbosity (R04, R05)

- Problem signal: experts report clusters being “obscure” and titles too long/AI-toned.
- Refinement: reinforce provenance cues (e.g., consistent AI badge, “AI-generated” label on cluster cards, clearer legend entry for clusters as AI thematic grouping).

Govern salience cues more defensively (R16; also impacts R11)

- Problem signal: node size repeatedly cited as the most bias-prone cue, even when explained.
- Refinement options: compress the size scale (cap extremes), add an optional toggle to neutralise size (uniform sizing), or add a

second, explicitly “non-normative” encoding (e.g., show degree on-demand rather than continuously).

Strengthen contestation as structured critique, not only a click (R17)

- Problem signal: request for comments/notes attached to contestation; current implementation confirms contestation but does not capture rationale/audit trail.
- Refinement: extend contest action with a short text rationale or selectable error categories (wrong grouping / misleading bridge / missing nuance), and (in future integration) store contestations as objects that can be reviewed and acted upon.

Re-scope “reflection scaffolds” to match what the interface truly does (R06)

- Problem signal: In the current platform, reflection is implemented downstream as analysis-time personal notes (Q16 high, but with requests for saving/sharing and explicit privacy framing).
- Refinement: Specify that personal notes are private and isolated on the local session. make it more explicit that the export view includes also the notes and drawings.

Requirements excluded from the final validated set for the current artefact

To avoid overstating what the current implementation can credibly guarantee, the expert validation supports defining two tiers:

- Tier 1 - validated as implemented in Il Corollario: Strongly supported by high ratings and consistent with the implemented interaction grammar: R01, R02, R03, R05, R07, R08, R09, R10, R13, R16.
- Tier 2 - only partially instantiated / should remain “future or system-level requirements”

R04 (provenance marking): present but unevenly perceived; needs clearer provenance cues, especially for clusters.

R06 (upstream reflection scaffolds): implemented as downstream notes; should be rephrased for the KG-only scope.

R11 (defensive design against manipulation): partially addressed via alternative views, but lacks stronger defensive layers (and received the lowest mean score).

R12 (common ground mediation): approximated via shared keywords, but no workflow for negotiated synthesis/revision.

R15 (parameters/thresholds & re-aggregation controls): currently reduced to “inspectable criteria” via tooltips when the cursor is above the selector; no actual controls/sensitivity exploration; strong discoverability dependency.

R17 (revision workflow): contest actions exist, but without an auditable, system-level revision loop.

Finally, R14 (disclosure of normative operationalisations) is part of the Chapter 5 catalogue but is not directly evaluated by a dedicated survey item and is not explicitly operationalised as a distinct interface mechanism in the current KG artefact.

Survey Q#	Survey task	Survey rating statement	Req. ID	Requirement title	Requirement descriptor	ORBIS filtered req. ID(s)	Survey scores (1–10)
Q01	Going through the tutorial at the beginning of the experience.	The step-by-step tutorial helps me understand how to navigate the discussion structure before entering it.	R01	Progressive Disclosure & Narrative Sequencing	Designers must structure complex information using narrative sequencing (e.g., "scrollytelling," "Martini Glass" structure). Present context and conceptual framing before technical details or visualizations to scaffold understanding and prevent cognitive overload.	DAV.6, MA.2, MA.6, MPD.1, UIE.4	9.36 (SD 0.81); range 8–10
Q02	Using the depth-filters to reduce complexity on-demand.	Information is presented in layers (basic <> advanced) so both non-experts and experts could engage without overload.	R02	Audience-Adaptive Complexity	The complexity of visualizations and explanations must be adapted to the target audience's domain knowledge and literacy levels. Designers should account for different epistemic profiles (e.g., experts vs. lay citizens) and education levels, which moderate persuasion and comprehension.	DAV.6, UIE.4	8.45 (SD 1.37); range 6–10
Q03	Use the Legend to interpret node types (Subject, Position, In Favor, Against, Cluster Area, Keyword) and link styles.	The visualization form and how information is coded with colored shapes (knowledge graph + clusters + keywords) makes it easy understanding debate structure.	R08	Legible & Familiar Visualization Forms	Prioritize domain-legible and accessible visualization types (e.g., simple histograms, heatmaps, traffic-light indicators) over complex forms (e.g., large network graphs) that may intimidate non-expert participants. Ensure visual metaphors align with users' mental models.	DAV.6, UIE.4	9.27 (SD 1.1); range 7–10
Q04	Click a Position, then see Arguments connected; use the graph connections and details to verify relations	High-level elements (e.g., clusters/ keywords) are anchored to underlying arguments through clear interactive links, making verification straightforward.	R03	Cross-Modal Evidential Anchoring	Textual claims (narratives, summaries, AI insights) must be explicitly linked to visual evidence. Use interactive mechanisms like hover-to-highlight or hyperlinks to bind high-level insights back to specific source data (e.g., original citizen proposals/comments) to support verification and reduce text dominance.	DAV.2, RS.2, RS.5, UIE.1	8.18 (SD 1.78); range 5–10
Q05	See Solid Line vs Dashed Line (Legend and Graph); then open a Keyword or Cluster node in Details.	The interface clearly distinguishes human-authored content from AI-generated elements/links—so provenance is easy to judge.	R04	Explicit Provenance & Authorship Marking	The interface must visually distinguish between human-authored content and AI-generated syntheses (e.g., using specific icons or distinct visual styles). Designers must employ provenance rhetoric (citations, methodology notes) to maintain credibility and allow users to assess the source of the information.	RS.5, UIE.1	7.55 (SD 2.5); range 2–10
Q06	Open a Cluster or Keyword, expand AI Role in processing, and read the confidence framing.	Deliberation KG communicates AI limitations/uncertainty in a visible and non-overconfident way.	R05	"Seamful" Design & Uncertainty Communication	Avoid presenting AI outputs or data stories as flawless or neutral truths. Interfaces should practice seamful design by visibly exposing system limitations, uncertainties, and data gaps (e.g., model reality vs. lived reality) to encourage epistemic vigilance and prevent over-reliance.	DAV.2, RS.2	8.1 (SD 1.8); range 4–10
Q07	Use Suggested Views to activate specific perspectives: Most Contested; Shared Keywords, Lone Voices.	Filtering views help me compare perspectives and eventually test hypotheses, rather than accepting a single given framing.	R07	Interactive "What-If" & Comparison Tools	Move beyond static explanations by providing what-if simulations, comparative views (contrastive explanations), and filtering capabilities. This supports human-centered reasoning by allowing users to test hypotheses, see consequences of changes, and compare diverse scenarios.	CF.1, DAV.4	7.8 (SD 1.7); range 5–10
Q08	Use the Timeline to replay the debate's evolution.	I can see the evolution of the debate, revisit earlier states and reproduce a specific view of the debate (supporting traceability and contestation over time).	R09	Support for Traceability, Contestability & History	The system must allow users to trace how AI outputs were produced (from source data to clustering/summaries) and to navigate the history of the discussion. Provide interaction primitives such as Extract (save/ share evidence) and Replay (undo/view history) so users can contest interpretations, reproduce views, and revisit earlier states.	MPD.1, MPD.4	8.2 (SD 2.5); range 2–10
Q09	Find how the system decides things like 'Most contested', 'Lone voices', 'shared keywords '.	Key parameters behind visibility (e.g., what counts as 'contested' or concept bridges) are inspectable and understandable.	R15 + (supports R16)	Legibility of Parameters, Thresholds & Re-Aggregation Controls	Users should be able to inspect the key parameters, thresholds, and metrics that drive aggregation and visibility (e.g., cluster granularity, salience scoring, filtering rules). Provide robust defaults plus safe controls or sensitivity views that reveal how outcomes change under different settings.	—	7.9 (SD 2.5); range 1–10
Q10	Pick a Position and examine both In Favor and Against arguments; optionally activate Lone voices filter (suggested view section).	The interface makes disagreement and minority viewpoints visible—as to say: it doesn't flatten diversity into false consensus.	R10	Representation of Plurality & Conflict	AI synthesis tools should not artificially smooth over disagreements to create a false consensus. Design representations (e.g., clusters, argument maps) that make minority viewpoints, distinct issue communities, and conflict structures visible and legitimate.	CF.1, DAV.1, DAV.4, MA.4	8.36 (SD 1.3); range 6–10

Survey Q#	Survey task	Survey rating statement	Req. ID	Requirement title	Requirement descriptor	ORBIS filtered req. ID(s)	Survey scores (1–10)
Q11	Use at least two alternative views/filters and see whether they reveal different interpretations of the same debate.	The interface supports critical reading by encouraging cross-checking: these alternative views help reduce framing/manipulation risk.	R11 + (relates to R07, R16)	Defensive Design Against Manipulation	Anticipate that narrative and rhetorical choices (omission, emphasis, framing) can manipulate interpretation. Implement defensive design strategies (e.g., exposing alternative framings, rigorous fact-checking layers) to protect against misinformation and dark patterns in data storytelling.	—	7.45 (SD 1.9); range 4–10
Q12	Look at Node size and read the Tutorial explanation about size; consider how information encoding might shape interpretation.	After reading the explanation, I understand that visual emphasis (such as node size) is a design choice and not a judgment about truth or value.	R16	Editorial & Rhetorical Transparency Cues	Make editorial choices legible (ordering, emphasis, omissions, visual rhetoric) and surface how these choices shape interpretation. Provide alternative framings or counter-views to reduce the risk of persuasive bias being mistaken for neutrality.	MA.6	8.3 (SD 1.56); range 6–10
Q13	Use Shared keywords (concept bridges) to find where opposing arguments touch the same concept.	The interface helps identify potential common ground and supports a workflow for critically assessing it (not just accepting AI mediation).	R12	Facilitation of "Common Ground" via Mediation	Use AI as a mediator to draft synthesized statements that represent the group's common ground. Design a workflow that allows participants to critique and revise these AI drafts, ensuring the final narrative is a negotiated consensus rather than an algorithmic imposition.	—	8.1 (SD 1.9); range 4–10
Q14	Open "From data to knowledge" and review how clustering and keyword extraction are explained, including their limitations.	Deliberation KG clearly discloses how its analytical transformations (clustering and keyword extraction) shape what I see, making it clear that these results are interpretative processes rather than neutral facts.	R13	End-to-End Transformation Ledger	The interface must expose the main transformations applied to deliberative input (selection, filtering, clustering, summarization, ranking), including what is excluded and why. Provide an end-to-end transformation ledger that supports accountability and interpretation of AI-mediated outputs.	MPD.4	8.1 (SD 1.45); range 6–10
Q15	Open an Argument or Keyword and use the Contest AI option.	The interface provides clear, in-situ mechanisms to challenge AI outputs with feedback, and becomes clear that feedback is captured for revision.	R17	In-Situ Feedback, Disagreement Capture & Revision Workflow	Provide mechanisms for participants to challenge AI outputs (clusters, summaries, labels) with structured feedback (e.g., wrong grouping, missing viewpoint, distorted summary). Preserve an audit trail of revisions and rationales, and support critique-driven iteration on mediated statements.	—	8.3 (SD 1.35); range 6–10
Q16	Open Personal Notes, add a note (and optionally a drawing) while inspecting part of the debate.	Deliberation KG supports reflective engagement (e.g., personal annotation space) that helps users annotate insights or patterns while analysing the discussion.	R06	Upstream Reflection Scaffolds	Incorporate reflection nudges (e.g., persona prompts, storytelling prompts) during the drafting/input phase. These communicative scaffolds intervene upstream to improve the quality, diversity, and constructiveness of user contributions before they enter the public discussion.	—	8.45 (SD 1.6); range 5–10
Q17	Think about the whole experience after exploration.	After a short exploration, I can use Deliberation KG fluently enough to analyse and make sense of a deliberative discussion.	—	— (global evaluation item)	—	—	8.3 (SD 1.3); range 6–10
Q18	Consider plurality + traceability + contestability together.	Overall, Deliberation KG supports democratically adequate interpretation (plurality, transparency, contestability) rather than persuasion.	—	— (global evaluation item)	—	—	8.2 (SD 1.9); range 5–10
Q19	Consider real-world usage as a complement to linear threads.	I would recommend Deliberation KG as a complementary view to support facilitation and public understanding in real deliberative processes.	—	— (global evaluation item)	—	—	8.8 (SD 1.3); range 7–10

Tab. 13 Expert validation survey's questions mapped to SLR and ORBIS requirement with evalutations scores

8.0
DISCUSSION

This thesis started from a tension introduced in Chapter 1: AI is increasingly used to make large-scale deliberation manageable, yet those same pipeline functions can widen an interpretability gap by making it harder for heterogeneous publics to understand, verify, and contest how collective representations are produced. The work therefore reframed explainability in civic deliberation as a communication-design problem: the critical unit is the mediation pipeline as it is encountered through public-facing representations, not only the model.

This chapter discusses what the validation actually supports, what the thesis contributes across HCXAI, deliberation, and communication design, and how the claims are bounded to avoid over-claiming beyond the evaluated evidence.

8.1 Research Contributions to knowledge

Reframing civic explainability from model transparency to pipeline legibility (RQ0, RQ1)

Claim. Civic explainability in AI-enhanced deliberation should be treated as transformation transparency: publics need legible accounts of how discourse is selected, clustered, annotated, and rendered salient, with inspectable evidence traces, provenance cues, and seamful disclosure of limits (RQ1).

Where it is demonstrated. The reframing is articulated through the SLR synthesis and operationalized as requirements and principles (Chapter 5), then instantiated as concrete disclosure and inspection mechanisms in Il Corollario (Chapter 6), and tested through user tasks targeting AI-construct comprehension and verification behaviors (Chapter 7).

Evidence. Expert validation indicates that the end-to-end disclosure of analytical transformations ("From data to knowledge") is perceived as making clustering and keyword extraction legible as interpretive processes (Q14 mean 8.3/10, SD 1.3). User testing further shows that participants can generally act on the pipeline-legibility premise (overall task completion 96.15% across 130 observations; mean confidence 4.59/5), while the most demanding moments concentrate precisely where transformation transparency becomes operational (keyword/bridge verification in text, and cross-ecosystem evidence bridging).

Boundary. This evidence does not prove that the underlying AI mediation is "correct," nor that participants can always interpret it appropriately under real-world time pressure; it supports that the interface can make mediation inspectable when cues are encountered. Discoverability remains a structural condition: if disclosure entry points are missed, the same system can revert to "AI as authority" misreadings despite being seamfully designed.

An evidence-anchored requirements design framework bridging HCXAI, data storytelling and deliberation (RQ0, RQ2, RQ3)

Claim. The thesis contributes a requirements-driven communication-design framework that translates cross-field evidence into auditable design criteria for civic explainability, supporting plurality-preserving sensemaking (RQ2) and an evaluation stance beyond trust (RQ3), anchored to explicit requirements rather than generic "transparency" claims.

Where it is demonstrated. The framework is derived from the SLR's four meta-themes, formalized as requirements (Chapter 5), aligned with ORBIS constraints, and then used as the evaluative backbone in Chapter 7 via a direct mapping from expert survey items to requirement IDs.

Evidence. The expert validation shows overall strong perceived democratic adequacy for the requirement set as instantiated in the artifact (mean 8.36/10 across 19 items), with particularly high ratings for entry scaffolding and legible visual grammar, tutorial onboarding (R01; Q01 mean 9.4/10) and legend/visual grammar comprehension (R08; Q03 mean 9.4/10). At the same time, the ratings surface exactly where plurality-preserving sensemaking and "evaluation beyond trust" become fragile in practice: cross-checking/defensive reading via alternative views (R11; mean 7.5/10), provenance clarity (R04; mean 7.7/10), and parameter/criteria inspectability (R15; mean 7.9/10) show lower means and higher dispersion, indicating that "being present" is not equivalent to "being reliably usable as a civic safeguard."

Boundary. The framework is validated here as a design-and-evaluation scaffold within the ORBIS/BCause context; it is not evidence that these requirements are universally sufficient for democratic legitimacy across domains or populations. Requirements that were not directly testable (or not explicitly operationalized as distinct mechanisms in the current artifact) must remain outside the validated claim set.

A validated design artifact demonstrating "communication design as mediation" in a real civic-tech ecosystem (RQ0 as artifact-level proof)

Claim. Il Corollario demonstrates that coupling network visualization with narrative and interaction scaffolds can support orientation-first sensemaking of AI-enhanced deliberations while keeping AI-mediated layers contestable and non-authoritative.

Where it is demonstrated. The claim is instantiated through the artifact's tutorial/legend "interpretive contract," typed deliberation roles (subject, positions, arguments), evidence-oriented inspection (details-on-demand), disclosure panels, and in-situ contestability actions (Chapter 6), and tested via task-based user sessions and expert requirement probes (Chapter 7).

Evidence. User interviews indicate that a dense deliberation map can be learnable and usable when progressive disclosure is treated as design infrastructure: task success reached 96.15% with strong confidence (4.59/5), and perceived usability was high (SUS mean 82.5/100).

Expert validation further supports the artifact as a complementary view for real deliberation/facilitation contexts (recommendation item mean 9.0/10) and as supporting democratically adequate interpretation "rather than persuasion" (Q18 mean 8.3/10).

Boundary. The artifact is validated as a prototype under ORBIS constraints and controlled sessions; this does not establish long-term effects on deliberative outcomes, minority influence, or institutional accountability practices. Moreover, the most demanding user tasks (e.g., cross-ecosystem verification T12 and "AI elements + contestability action" T13) show that democratic adequacy depends disproportionately on whether verification and challenge pathways are encountered and understood at the moment they are needed.

Requirement-driven validation as a replicable evaluation pattern for civic explainability.

Claim. The thesis contributes an evaluation structure that separates usability/learnability from requirement-level democratic adequacy by validating an artifact against an explicit requirement catalogue rather than relying on generic UX satisfaction or trust measures.

Where it is demonstrated. The two-moment strategy (user task-based sessions + expert requirement mapping) is defined and executed in Chapter 7, explicitly linking task breakdowns to "meaning-critical" requirements (evidence bridging, AI seams, contestability) and producing a requirement coverage table with per-item statistics.

Evidence. The user study yields both performance signals (errors/hints clustered on verification and agency tasks) and perceived usability (SUS) to contextualize friction, while expert ratings directly identify priority improvement zones (R11, R04, R15, R07, R09) instead of flattening evaluation into a single "overall satisfaction" metric.

Boundary. The evaluation design supports requirement plausibility and breakdown identification, not population-level generalization. With n=10 users and n=11 experts, results are sufficient for discovering major interpretive failure modes and for bounding claims, but not for establishing stable effects across publics or institutional settings.

8.2 Communication Design as Mediation in Explainable AI for Civic Deliberation

This thesis treats mediation as the representational and interactional conditions that determine whether AI-assisted deliberation remains interpretable and contestable for heterogeneous publics. In Il Corollario, mediation is enacted through three coupled instruments, narrative scaffolding (tutorial + disclosure), visualization grammar (typed nodes, legible encodings), and interaction scripting (progressive disclosure, evidence inspection, comparison views, contestability). Validation supports this as a viable interface-level strategy: onboarding and visual grammar are the strongest enablers of interpretive access, while the fragilities reveal where mediation can fail even when mechanisms exist.

8.2.1 How gaps were mitigated, met and where they remain partial

From "showing connectivity" to deliberative legibility (gap: representation mismatch)

A key ORBIS baseline problem was that the initial KG view (Grafana) behaved as a technical output: nodes were visually homogeneous and deliberative roles were hard to read, increasing interaction costs and fragmenting interpretation. Il Corollario mitigates this by making role

legibility the primary communicative goal: Human-contribution nodes, distinct deliberative functions (subject, positions, arguments) and a stable "interpretative contract" supported by tutorial onboarding and a legend/ how to read grammar. This directly instantiates the thesis claim that explainability failures often occur "after the model", in the representational layer.

Making verification a default reading practice (gap: weak evidence pathways)

The framework's cross-modal evidential anchoring requirement (R03) is addressed through details-on-demand payloads and inspection loops that keep links to source contributions available during exploration. Validation supports this direction but also shows the boundary condition: the most demanding user tasks were those that required cross-ecosystem verification (e.g., locating discussion content inside the graph) and keyword verification in text, indicating that verification is not only a feature but a "friction budget" problem. This is a concrete contribution to civic explainability: even when traceability exists, democratic adequacy depends on making verification low-cost enough to be routinely practiced.

Seamful AI and the legitimacy of uncertainty (gap: "AI as authority" misreadings)

The framework positions seamful design and limitation communication as central to civic settings (R05), and expert feedback indicates that the artifact's "seamful framing" is generally recognized (e.g., uncertainty/ limitations and "From data to knowledge" disclosures). However, expert and user feedback converge on a critical mediation insight: adequacy depends on discoverability at the moment of interpretation. When disclosures are optional and easily missed, users may still over-read AI clusters/keywords as objective structure. The design implication is structural: seamfulness must be "ambient," not just available.

Contestability as epistemic agency, not just a button (gap: procedural affordances without interpretive access)

A core claim is that contestability presupposes interpretive access: people can only challenge what they can understand as constructed, bounded, and revisable. Il Corollario operationalizes contestability through in-situ challenge affordances; experts explicitly valued this as an agency instrument, not only as transparency disclosure. Yet the validation also clarifies what remains missing at system level: a structured revision

workflow with rationale capture and an auditable loop (R17) was only partially instantiated, and experts requested turning contestation into structured critique with stored rationales. This distinction is important for the thesis' knowledge claim: interface-level contestability can be demonstrated, but full democratic recourse requires integration with institutional processes of review and revision.

Defensive design against framing and salience bias (gap: rhetorical side-effects of visualization)

The evaluation highlights a persistent risk aligned with the literature on rhetorical effects: node size and other salience cues can steer attention and be misread as normative judgment even when explained. Experts identified salience governance as a priority improvement zone, motivating defensive options such as neutral modes or compressing size extremes. This finding extends the thesis argument: communication design does not merely "clarify" AI; it also introduces its own rhetorical force, so civic explainability must include explicit controls and safeguards for interpretive bias.

Validated vs future/system-level requirements

Expert validation supports a two-tier classification to avoid overstating democratic guarantees. Tier 1 requirements are strongly supported as implemented (R01, R02, R03, R05, R07, R08, R09, R10, R13, R16). Tier 2 requirements remain partial or out of scope for the current artifact (e.g., provenance marking clarity, defensive design, parameter controls, common-ground mediation, revision workflow). The second tier works as a blueprint of what could be implemented in the platform in the future and also what needs to be redesigned or reinforced from the already existing set of features implemented in Il Corollario.

8.3 Methodological Contributions and Limits of the research

8.3.1 Methodological contributions

The thesis contributes a traceable method chain: (i) a systematic review designed for cross-field integration (not single-discipline completeness), (ii) meta-themes aligned with RQs, (iii) evidence-anchored requirements

Chapter // 8

to build a design framework, and (iv) an artifact that operationalizes requirements as design mechanisms.

Requirement-driven validation beyond generic UX evaluation: The evaluation strategy explicitly separates usability/learnability (user tasks + SUS) from requirement-level democratic adequacy (expert validation mapped to the Chapter 5 requirement set).

8.3.2 Limits

Scope of deployment and ecosystem integration

Il Corollario is validated as a prototype/artefact within ORBIS constraints; several “full legitimacy” conditions (e.g., auditable revision loops, governance of disputes, negotiated synthesis) require deeper platform integration and institutional adoption to be properly evaluated.

Sample size and evaluation horizon

User (n=10) and expert (n=11) evaluations are sufficient for identifying major breakdowns and for requirement-level plausibility, but they cannot establish population-level effects or long-term impacts on deliberative outcomes. The thesis therefore supports claims about interface-level interpretability and perceived democratic adequacy, not about macro-level political legitimacy.

Discoverability as a structural validity risk

A recurrent limitation emerging from results is that many “good” mechanisms (provenance cues, disclosures, parameter explanations) can fail if not encountered at the right moment. This implies that future evaluation should treat discoverability and timing of cues as first-order variables, not as minor UX polish.

ORBIS project constrains

The initial datasets exported from Grafana that were provided as a baseline to start developing the platform influenced some of the design choices of the platform. Some of the features implemented to respond to Requisites were limited from the data structure, designed around them instead of following what the best deliberation function would be, and these limitations emerged in the user test. For example the keyword extraction is made only on arguments and not positions, and the clustering of arguments is limited to the user pro/con choice made when creating the contribution on BCause. A possible future direction should take in consideration how data needs to be collected and structured in order to best serve the visualization platform that is going to give back to the user the data itself.

9.0
CONCLUSIONS

Artificial Intelligence is progressing at a pace that makes it increasingly hard—especially for non-specialists—to keep up with technical developments and to anticipate their implications for citizens. In democratic deliberation, AI-mediated pipelines can unlock real benefits: they can help make participation manageable at scale, support synthesis and orientation, and reduce the practical burden of navigating large volumes of public input. At the same time, the literature reviewed in this thesis shows that these interventions introduce new ethical and democratic risks: normative operationalizations can privilege some communication styles, pipeline stages can hide exclusions and distortions, and "helpful" representations can be misread as neutral truths unless their construction is made visible and contestable.

Against this background, the thesis argued for a shift from model-centric explainability toward the intelligibility of mediation as it is encountered through public-facing interfaces.
In other words, the core requirement is not that citizens understand the internal mechanics of an AI model, but that they can understand what transformations were applied to deliberative data (selection, aggregation, clustering, summarization, ranking), what those outputs legitimately mean, and how to verify and challenge them when they are misleading or incomplete. This reframing positions communication design as a mediating layer: a discipline that can make transformation processes explicit in a clear, comprehensible, and sincere way, without falling into two symmetric failures that are equally ineffective in civic contexts: explaining everything (overload and technicism) or explaining nothing (complete opacity).

Within ORBIS and the BCause reporting ecosystem, this thesis operationalized that claim through two concrete outcomes: (i) an evidence-based requirements framework grounded in a systematic literature review across HCXAI, visualization/storytelling, and deliberation, and (ii) Il Corollario as a research-through-design artifact that instantiates those requirements as a typed, provenance-aware, and contestable knowledge-graph interface for mapping deliberative discussions.

9.1 Implications for Civic Tech and AI Governance

First, the thesis reinforces a governance implication that is often underestimated: when AI mediates participation, explainability is not a user-experience enhancement, it becomes part of the public interface of democracy and therefore a condition for legitimate participation.

If civic-tech systems rely on clustering, summarization, ranking, or civility/rationality proxies, then governance must require public-facing legibility of what the system treats as relevant, what it excludes, and how those choices shape visibility and interpretation. In practical terms, this suggests that integration, evaluation, and deployment of AI in civic contexts should include transformation-transparency and contestability requirements as first-order criteria, not optional documentation.

Second, the results suggest a shift in what "responsible AI" means at interface level. The most fragile points are not only model errors but communicative failures: provenance ambiguity (human vs AI content), parameter opacity behind salience cues, and low discoverability of challenge mechanisms. Governance implications follow directly: if citizens cannot reliably tell what is computed versus what is authored, or cannot inspect why certain elements are emphasized, the system risks performing persuasion-by-interface even when it intends to support understanding.

Third, the thesis clarifies the role of interdisciplinary collaboration. Many challenges remain highly technical (pipeline reliability, bias, lossless conversion limits), but communication design contributes a distinct and necessary competence: making complex mediation honest and usable under cognitive constraints, staging what needs to be understood, and designing for verification and disagreement as normal civic practices.

9.2 Future Research Directions

Close the loop from "contestation as signal" to "contestation as revision workflow."

Il Corollario currently demonstrates in-situ contestability as an agency mechanism, but it doesn't dive into how the feedback needs to be structured and how it will influence the pipeline. Future work should therefore integrate structured disagreement capture with an auditable revision process (who reviewed, what changed, why), aligning interface-level contestation with institutional accountability.

Strengthen provenance and parameter inspectability

Expert evaluation indicates that provenance clarity and inspectability of criteria behind "contested / lone voice / shared keywords" are priority improvement zones, and that adequacy depends on whether these mechanisms feel auditable in practice, not just present in principle. Future iterations should make provenance distinctions and lens criteria more accessible (visible in default reading), not only through optional panels.

Chapter // 9

Defensive design against rhetorical effects of visualization

Both user and expert feedback converge on the risk that salience cues (node size, emphasis, suggested views) can steer attention and be misread as normative judgment. A concrete research direction is to develop and test "neutral reading modes" and staged reveal strategies, treating rhetorical transparency as a measurable design variable rather than a qualitative aspiration.

Generalize the ingestion and mapping pipeline

External feedback from within ORBIS suggests that the approach could be valuable beyond BCause reports. This motivates a research direction: building modular ingestion adapters to apply the same IBIS-like typed mapping to other high-volume public discussions (e.g., large forum threads) in order to stress-test the limits of conversion, bias introduced by structuring, and the robustness of transformation-transparency disclosures across contexts. The goal here would be methodological and evaluative (generalizability of the framework and artifact), not productization.

Evaluation in real facilitation settings

The current validation demonstrates legibility and democratic adequacy in controlled sessions. Future work should test whether verification and contestation behaviors persist over time in real deliberative processes: whether facilitators and participants actually use traceability pathways, how often contestation is exercised, and whether the interface changes outcomes such as perceived legitimacy, minority visibility, and accountability practices.

Finally, Il Corollario is released and will remain an open-source platform, with the intent that the work can be extended, audited, and adapted by others. In a field moving as fast as AI, this openness is not only a dissemination choice: it is coherent with the thesis' core claim that democratic adequacy depends on inspectability, interdisciplinary critique, and collective iteration rather than closed, authoritative representations.

BIBLIOGRAFY

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alexandre, I. (2016). Promoting insight: A Case Study of How to Incorporate Interaction in Existing Data Visualizations. In E. Banissi, M. Bannatyne, F. Bouali, R. Burkhard, J. Counsell, U. Cvek, M. Eppler, G. Grinstein, W. Huang, S. Kernbach, C. Lin, F. Lin, F. Marchese, C. Pun, M. Sarfraz, M. Trutschl, A. Ursyn, G. Venturini, T. Wyeld, & J. Zhang (Eds), PROCEEDINGS 2016 20TH INTERNATIONAL CONFERENCE INFORMATION VISUALISATION IV 2016 (pp. 203–208). ViS; BUILTVIZ; GMAI; MEDi VIS; CGiV; D Art. https://doi.org/10.1109/IV.2016.15

Bachiller, S., Quijano-Sanchez, L., & Cantador, I. (2021a). A flexible and lightweight interactive data mining tool to visualize and analyze digital citizen participation content. 36TH ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, SAC 2021, 413–416. https://doi.org/10.1145/3412841.3442081

Bachiller, S., Quijano-Sanchez, L., & Cantador, I. (2021b). A flexible and lightweight interactive data mining tool to visualize and analyze digital citizen participation content. In 36TH ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, SAC 2021 (pp. 413–416). ASSOC COMPUTING MACHINERY. https://doi.org/10.1145/3412841.3442081

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barvosa, E. (2019). Mapping and Measuring Deliberation: Towards a New Deliberative Quality. By André Bächtiger and John Parkinson. Oxford: Oxford University Press, 2019. 208p. $78.00 cloth. Perspectives on Politics, 17(4), 1170–1172. https://doi.org/10.1017/S1537592719003323

Bhatia, A., & Sukthankar, G. (2025). Using LLMs to Structure and Visualize Policy Discourse. In W. Yin, J. J. Ahn, R. Zhang, L. Huang, R. Hadfi, T. Ito, S. Ohnuma, & S. Shiramatsu (Eds), Artificial Intelligence for Research and Democracy (pp. 69–76). Springer Nature. https://doi.org/10.1007/978-981-97-9536-9_5

Bobek, S., Korycinska, P., Krakowska, M., Mozolewski, M., Rak, D., Zych, M., Wojcik, M., & Nalepa, G. J. (2025). User-centric evaluation of explainability of AI with and for humans: A comprehensive empirical study. In INTERNATIONAL JOURNAL OF HUMAN-COMPUTER STUDIES (Vol. 205). ACADEMIC PRESS LTD-ELSEVIER SCIENCE LTD. https://doi.org/10.1016/j.ijhcs.2025.103625

Boehnert, J. (2015). The Politics of Data Visualisation. Discover Society, 23. http://discoversociety.org/2015/08/03/viewpoint-the-politics-of-data-visualisation/?utm_content=bufferd1e0e&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Borges, M. G., Correa, C. M., & Silveira, M. S. (2025). Decoupling narrative visualizations into components: Helping designers to tell data-driven stories. INFORMATION VISUALIZATION, 24(2), 135–149. https://doi.org/10.1177/14738716241284599

Buhmann, A., & Fieseler, C. (2023). Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence. BUSINESS ETHICS QUARTERLY, 33(1), 146–179. https://doi.org/10.1017/beq.2021.42

Carstens, J. A., & Friess, D. (2024). AI Within Online Discussions: Rational, Civil, Privileged? MINDS AND MACHINES, 34(2). https://doi.org/10.1007/s11023-024-09658-0

Chambers, M., & Denham, B. (2025). Measuring the Persuasive Effects of Narrative Visualizations. 2025 IEEE International Professional Communication Conference (ProComm), 296–304. https://doi.org/10.1109/ProComm64814.2025.00063

Dörk, M., Feng, P., Collins, C., & Carpendale, S. (2013). Critical InfoVis: Exploring the politics of visualization. CHI '13 Extended Abstracts on Human Factors in Computing Systems, CHI EA '13, 2189–2198. https://doi.org/10.1145/2468356.2468739

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning (arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2024). The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In PROCEEDINGS OF THE 2024 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYTEMS (CHI 2024). ASSOC COMPUTING MACHINERY. https://doi.org/10.1145/3613904.3642474

Ehsan, U., & Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In C. Stephanidis, M. Kurosu, H. Degen, & L. Reinerman-Jones (Eds), HCI International 2020—Late Breaking Papers: Multimodality and Intelligence (pp. 449–466). Springer International Publishing. https://doi.org/10.1007/978-3-030-60117-1_33

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks (arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Grancea, I., & Tutui, V. (2025). Scaling Up Good Listening: An Assessment Framework for AI-Powered Mass Deliberation Models. MEDIA AND COMMUNICATION, 13. https://doi.org/10.17645/mac.9961

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey Of Methods For Explaining Black Box Models (arXiv:1802.01933). arXiv. https://doi.org/10.48550/arXiv.1802.01933

Gutmann, A., & Thompson, D. (2004). Why Deliberative Democracy? Princeton University Press.

Haugeland, J. (1985). Artificial Intelligence: The Very Idea. MIT Press.

He, Y., Xu, K., Cao, S., Shi, Y., Chen, Q., & Cao, N. (2025). Leveraging Foundation Models for Crafting Narrative Visualization: A Survey. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 31(10), 9303–9323. https://doi.org/10.1109/TVCG.2025.3542504

Herrera, F. (2025). Reflections and attentiveness on eXplainable Artificial Intelligence (XAI). The journey ahead from criticisms to human–AI collaboration. Information Fusion, 121, 103133. https://doi.org/10.1016/j.inffus.2025.103133

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network (arXiv:1503.02531). arXiv. https://doi.org/10.48550/arXiv.1503.02531

Jessica Hullman & Nick Diakopoulos. (2011). Visualization Rhetoric: Framing Effects in Narrative Visualization. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2231–2240. https://doi.org/10.1109/TVCG.2011.255

Kempeneer, S., Wolswinkel, J., & van Maanen, G. (2025). Beyond the Portal: Increasing Citizen Understanding of OGD With Storytelling. INFORMATION POLITY, 30(1), 65–81. https://doi.org/10.1177/15701255241304651

Knuth, D. E. (1997). The art of computer programming, volume 1 (3rd ed.): Fundamental algorithms. Addison Wesley Longman Publishing Co., Inc.

Lipton, Z. C. (2018). The mythos of model interpretability. Commun. ACM, 61(10), 36–43. https://doi.org/10.1145/3233231

Lyu, Y., Lu, H., Lee, M. K., Schmitt, G., & Lim, B. Y. (2024). IF-City: Intelligible Fair City Planning to Measure, Explain and Mitigate Inequality. In IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS (Vol. 30, Issue 7, pp. 3749–3766). IEEE COMPUTER SOC. https://doi.org/10.1109/TVCG.2023.3239909

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ridley, M. (2025). Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper. In JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY (Vol. 76, Issues 1, SI, pp. 98–120). WILEY. https://doi.org/10.1002/asi.24889

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. In IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (Vol. 46, Issue 4, pp. 2104–2122). IEEE COMPUTER SOC. https://doi.org/10.1109/TPAMI.2023.3331846

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536. https://doi.org/10.1038/323533a0

Russell, S. J., & Norvig, P. (with Davis, E., & Edwards, D.). (2016). Artificial intelligence: A modern approach (Third edition, Global edition). Pearson.

Scharowski, N., Perrig, S. A. C., Svab, M., Opwis, K., & Bruhlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. In FRONTIERS IN COMPUTER SCIENCE (Vol. 5). FRONTIERS MEDIA SA. https://doi.org/10.3389/fcomp.2023.1151150

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. Neural Networks, 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Segel, E., & Heer, J. (2010). Narrative Visualization: Telling Stories with Data. IEEE Transactions on Visualization and Computer Graphics, 16(6), 1139–1148. https://doi.org/10.1109/TVCG.2010.179

Shao, H., Martinez-Maldonado, R., Echeverria, V., Yan, L., & Gasevic, D. (2024). Data Storytelling in Data Visualisation: Does it Enhance the Efficiency and Effectiveness of Information Retrieval and Insights Comprehension? In PROCEEDINGS OF THE 2024 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYTEMS (CHI 2024). ASSOC COMPUTING MACHINERY. https://doi.org/10.1145/3613904.3643022

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. Proceedings 1996 IEEE Symposium on Visual Languages, 336–343. https://doi.org/10.1109/VL.1996.545307

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. SCIENCE, 386(6719). https://doi.org/10.1126/science.adq2852

Tim Miller. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Ullmann, T. D., De Liddo, A., & Bachler, M. (2019). A Visualisation Dashboard for Contested Collective Intelligence. Learning Analytics to Improve Sensemaking of Group Discussion. In RIED-REVISTA IBEROAMERICANA DE EDUCACION A DISTANCIA (Vol. 22, Issue 1, p. 41+). ASOC IBEROAMERICANA EDUCACION SUPERIOR DISTANCIA. https://doi.org/10.5944/ried.22.1.22294

Vagnoni, S., & Palmirani, M. (2025). Fostering Deliberative Democracy in the Digital Age: An AI-Powered Platform for Enhanced Citizen Engagement in Legislative Processes. In L. Teran, J. Pincay, C. Vaca, & D. Riofrio (Eds), 2025 ELEVENTH INTERNATIONAL CONFERENCE ON EDEMOCRACY & EGOVERNMENT, ICEDEG (pp. 339–341). https://doi.org/10.1109/ICEDEG65568.2025.11081638

van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: Findings from 2 citizens' juries. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION, 28(10), 2128–2138. https://doi.org/10.1093/jamia/ocab127

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349

Weizenbaum, J. (1966). Eliza–a Computer Program for the Study of Natural Language Communication Between Man and Machine. Communications of the Acm, 9(1), 36–45.

Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation (pp. xii, 300). W. H. Freeman & Co.

Yeo, S., Lim, G., Gao, J., Zhang, W., & Perrault, S. T. (2024). Help Me Reflect: Leveraging Self-Reflection Interface Nudges to Enhance Deliberativeness on Online Deliberation Platforms. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, 1–32. https://doi.org/10.1145/3613904.3642530

Zeginis, D., Patsouras, C.-F., Promikyridis, R., Tambouris, E., & Tarabanis, O. (2026). Review of AI Features to Support Mass Deliberations. In I. Lindgren, M. Bolivar, M. Janssen, E. Loukis, F. Mureddu, P. Panagiotopoulos, G. Pereira, & E. Tambouris (Eds), ELECTRONIC GOVERNMENT, EGOV 2025 (Vol. 15944, pp. 336–350). https://doi.org/10.1007/978-3-032-01589-1_21

Zhang, A., Walker, O., Nguyen, K., Dai, J., Chen, A., & Lee, M. K. (2023). Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. Proc. ACM Hum.-Comput. Interact., 7(CSCW1). https://doi.org/10.1145/3579601

Zheng, C., & Ma, X. (2022). Evaluating the Effect of Enhanced Text-Visualization Integration on Combating Misinformation in Data Story. 2022 IEEE 15TH PACIFIC VISUALIZATION SYMPOSIUM (PACIFICVIS 2022), IEEE Pacific Visualization Symposium, 141–150. https://doi.org/10.1109/PacificVis53943.2022.00023

RINGRAZIAMENTI

A Ilaria, per la grandissima professionalità ed esperienza, per tutta la conoscenza che ha saputo trasmettermi e per la pazienza e supporto costante.

Al progetto ORBIS, senza il quale non sarebbe stato possibile tutto questo ed al suo Team con cui ho potuto confrontarmi ed imparare in questi mesi.

Alla mia famiglia, che mi ha permesso in questi anni di continuare sul mio percorso.

A tutti i miei amici che sono stati enormemente di supporto in questi mesi.