Enhancing Data Understandability. An Integrated Approach

Ilaria Mariani, Giacomo Garetto, Jennifer Edmond, Eleonora Lima, María Ruiz de Assín de los Santos, George Manias, Fabiana Fournier, and Lior Limonad

Abstract This chapter examines data understandability as a core epistemic condition for democracy and democratic deliberation, arguing that contemporary challenges stem less from data scarcity than from uneven capacities to interpret increasingly complex, AI-mediated information environments. Building on scholarship on visualisation literacy, epistemic inequalities, and the limits of model-centric XAI, the chapter reframes explainability as a matter of sensemaking rather than model disclosure. It theorises narrative scaffolding, critical interactivity, and participatory design as infrastructures capable of contributing to equitable interpretation in civic contexts. The analysis is grounded in two EU projects, KT4D and ORBIS, which operationalise these principles in practice and offer empirical evidence of how visualisation becomes

I. Mariani (✉)
Department of Design, Politecnico di Milano, Milan, Italy
e-mail: ilaria1.mariani@polimi.it

G. Garetto
School of Design, Politecnico di Milano, Milan, Italy
e-mail: giacomo.garetto@mail.polimi.it

J. Edmond · E. Lima
Trinity College Dublin, Dublin, Ireland
e-mail: edmondj@tcd.ie

E. Lima
e-mail: limae@tcd.ie

M. R. de Assín de los Santos
Cibervoluntarios, Madrid, Spain
e-mail: maria.ruizdeassin@cibervoluntarios.org

G. Manias
Department of Digital Systems, University of Piraeus, Piraeus, Greece
e-mail: gmanias@unipi.gr

F. Fournier · L. Limonad
IBM Haifa Research Lab, Haifa, Israel
e-mail: fabiana@il.ibm.com

L. Limonad
e-mail: liorli@il.ibm.com

I. Mariani et al. (eds.), *AI for Democracy*, Springer Series in Design and Innovation 66,
https://doi.org/10.1007/978-3-032-23948-8_3

explanation, narrative becomes interpretation, and co-creation becomes a method of democratic alignment.

Keywords Data narratives · Integrated approach · Data understandability · Public trust · Data legibility

As governments and public institutions increasingly rely on data to inform their decision-making, allocate resources, monitor services, and communicate policy outcomes, the challenge has shifted from data scarcity or quality to the difficulty of transforming abundant information into forms that diverse publics can actually understand and meaningfully use. In other words, the question is no longer how to generate data of quality, but how to render it in ways that are actionable, interpretable, and relevant for their audiences, often heterogeneous (Shao et al. 2024; Zhang et al. 2022).

In this light, it becomes crucial to acknowledge that—even amid unprecedented levels of data availability—comprehension remains unevenly distributed. Raw datasets, dashboards, and even standard charts often remain inaccessible. Raw datasets, dashboards, and even standard charts often remain inaccessible to citizens and stakeholders without high levels of data or visualization literacy. Extensive evidence shows that limited visualisation literacy constrains people's ability to draw accurate conclusions, recognise key insights, or assess uncertainty—ultimately affecting interpretation, recall, and trust (Börner et al. 2016; Boy et al. 2014; Firat et al. 2022; Morini et al. 2025). These gaps persist and span across demographic groups and educational levels. Empirical studies confirm indeed that individuals without formal analytical training struggle to make sense of dashboards or performance indicators (Echeverria et al. 2017; Pozdniakov et al. 2023). Maltese et al. (2015) document that even among higher-education students there are widespread difficulties in interpreting visual data, underscoring the need for design strategies that support comprehension rather than assume expertise.

In democratic contexts, these comprehension gaps are particularly problematic. They risk excluding those with fewer technical skills, reinforcing inequalities in who can participate in data-mediated decisions. Crucially, this makes the issue not simply cognitive but actually political. When only a portion of the population can meaningfully interpret evidence, democratic deliberation risks becoming dependent on specific elites, while those without or with lower levels of specialised literacy are left navigating policy debates with limited interpretive tools.

Compounding this, the dominance of numerical reasoning in public discourse—KPIs, indices, scores, dashboards—tends to further flatten multidimensional social phenomena into reductive metrics that obscure uncertainty, trade-offs, and causal complexity (Hullman and Diakopoulos 2011; Tufte and Graves-Morris 1983). As a result, both citizens and public-sector managers often struggle to connect datasets with lived experiences or institutional decision processes.

This landscape opens up a series of questions.

- *How can institutions move beyond simply providing data to ensuring that it becomes genuinely interpretable and meaningful for their publics?*
- *How can evidence be communicated in ways that allow people with different levels of literacy to participate on equal footing?*
- *What representational and communicative strategies can mitigate—rather than reinforce—epistemic inequalities?*
- *And, critically, which principles should guide the design of information environments so as to help transform complex evidence into forms that support, rather than hinder, public reasoning and democratic deliberation?*

This chapter is situated within the perimeter defined by these questions. It first outlines the current state of the art, then proposes a reframing of data narratives as epistemic infrastructures for democratic life. Finally, it examines how these principles are operationalised in KT4D and ORBIS—two projects that provide concrete spaces for experimentation, testing, and refinement of democratic data-understandability practices—thereby contributing valuable first-hand knowledge to this emerging field.

1 Data Understandability. The Power of Narrative

Data visualisation is broadly understood as the practice of transforming data into graphical or multimodal representations that allow audiences to detect patterns, trends, and relationships that might otherwise remain obscure. Yet, within public governance and deliberative democracy, visualisation serves the deeper function of shaping epistemic conditions under which citizens, stakeholders, and public managers form opinions and positions, weigh trade-offs, and participate in collective reasoning (Elstub and Escobar 2019; Goñi 2025; Vromen 2017). In this setting, data narrative and narrative visualisations—integrating charts, text, images, annotations, and sometimes interactive elements—are increasingly recognised as infrastructural component of how public problems are collectively understood, argued, and acted, becoming key mediators in this process that offer structured yet interpretable pathways through complex evidence landscapes (Amini et al. 2015; Segel and Heer 2010; Tong et al. 2018).

1.1 A Matter of Literacy

As previously noted, many publics possess limited visualisation literacy (Firat et al. 2022; Shao et al. 2024). Studies consistently show that even basic charts can be misread or misunderstood by non-experts (Börner et al. 2016; Maltese et al. 2015), and this challenge has only intensified as governments deploy dashboards, data portals, and real-time indicators. An example comes from the COVID-19 pandemic (Fareed et al. 2021; Zhang et al. 2023). The study of Zhang and colleagues (2023)

shows how COVID-19 dashboards became contested artefacts, sitting at the intersection of public demands for detailed, near real-time data and the constraints of existing privacy rules and institutional procedures. Designers were pressured to publish increasingly granular near real-time data—such as zip-code–level case numbers—while simultaneously having to comply with HIPAA-based de-identification thresholds, negotiate internal approval processes, and navigate leadership hesitations. These negotiations led to heterogeneous dashboard practices across states and agencies, which in turn produced confusion and, at times, distrust among the public.

Moreover, public reactions—ranging from anxiety about vaccine distribution to harmful misinterpretations of racial disparities—actively reshaped the design of these dashboards. Visualisations that unintentionally amplified fear or stigmatized communities were removed or reworked, and designers introduced supplementary narrative explanations to contextualise patterns and mitigate misreadings. The study shows that dashboards operated as boundary objects, mediating between designers, policymakers, and 'armchair epidemiologists', revealing how visualisations not only communicate data but also influence emotions, beliefs, and behaviours during crises. In this context, COVID-19 dashboards provide again a good example of how data visualisations can simultaneously inform and misinform, and how their design unavoidably involves normative judgments about what counts as responsible or 'correct' storytelling (Zhang et al. 2023).

These comprehension gaps and interpretive conflicts create epistemic asymmetries within democratic systems. While public institutions increasingly rely on data-driven tools to justify or design policy, many citizens lack the interpretive resources necessary to meaningfully analyse or contest the evidence on which decisions (should) rest. In this sense, data-driven participation risks becoming participation conditioned by data literacy rather than democratic entitlement—a form of structural inequality that emerges not from intentional exclusion, but from the growing complexity of the information environment itself.

1.2 Epistemic Inequalities and Representational Asymmetries

Scholars also warn about deeper infrastructural biases embedded within data systems. In line with critical communication design, data visualisations are not neutral artefacts but embed values, ideological assumptions, and decisions about what to foreground or omit (Boehnert 2016). Representational gaps are often linked to how data infrastructures operationalise sensitive categories (race, socioeconomic status, age, gender) and align with institutional risk-avoidance practices (e.g., restrictive privacy rules, operational silos). Concepts such as 'datawash' and 'darkdata' (Boehnert 2016, 2015) highlight how omissions—what remains unmeasured, unmodelled, or unvisualised—can reinforce dominant narratives while silencing marginalised groups. Such omissions are frequently entangled with institutional power asymmetries, shaping who appears in public evidence, how granularly, and under what interpretive frames.

Consequently, such asymmetries have direct implications for democratic deliberation: publics cannot contest what they cannot see, and they cannot evaluate a policy domain whose evidentiary landscape is structurally incomplete. Thus, epistemic inequality emerges not only from differential literacy but also from differences in whose realities become legible within public data systems. In the public sector, these blind spots can yield governance decisions that appear rigorously evidence-based but in fact rely on partial or selective representations of social reality.

Beyond issues of literacy and misinterpretation, an additional challenge stems from the cognitive effort required to make sense of complex information. Cognitive load theory distinguishes between intrinsic load, which arises from the inherent complexity of the material; extraneous load, produced by the way information is presented; and germane load, the mental effort devoted to constructing meaningful understanding (Sweller 1988; Sweller et al. 1998). Public-sector datasets often exhibit high intrinsic load due to their abstraction, multidimensional structure, and uncertain or heterogeneous variables. When visualisations are difficult to parse or require extensive searching and comparison, extraneous load increases, leaving fewer cognitive resources available for sensemaking. In deliberative settings, where audiences may vary widely in their familiarity with data or statistical conventions, these combined elements can hinder comprehension and reinforce epistemic asymmetries. Recognising how different forms of cognitive load operate provides a foundation for understanding why certain representational strategies—such as data narratives—are needed to support more accessible interpretation and consequently empower participation.

1.3 Integrated Approaches for Supporting Understanding

Narrative approaches to data understandability offer one response to these challenges. They combine visual, textual, and interpretive cues that guide audiences through complexity rather than overwhelm them. Narrative devices such as annotations, contextual framing, and stepwise sequencing have been shown to reduce cognitive load and support comprehension, particularly among low-literacy audiences (Amini et al. 2015; Segel and Heer 2010; Shao et al. 2024). Formats such as scrollytelling, data stories, or structured visual pathways (Mörth et al. 2023; Schneiders 2020) help highlight causality, show uncertainty, and connect abstract indicators to real-world conditions (Bach et al. 2018; Tong et al. 2018). Moreover, Shao et al. (2024) notes that the most effective narrative visualisations do not simply 'tell a story' but provide adaptive interpretive scaffolds: for example, automated highlights, dynamic transitions, or contextual call-outs that help viewers situate individual data points within broader patterns. These design features support sensemaking by offering just-in-time cues that reduce the cognitive effort required to interpret complex displays—particularly in interactive or multilayered dashboards. In civic contexts, such narrative visualisations can strengthen personal reflection, emotional resonance, and perceived relevance—key components of meaningful public engagement (Claes and Vande

Moere 2017). Similarly, work in political theory and deliberative democracy highlights storytelling as a core mechanism through which people make sense of complex public issues, negotiate meaning across diverse arenas, and ultimately articulate arguments (Boswell 2013; Bruner 1991). Recent work on AI-powered narrative building for digital governance further shows how algorithmic support—through Natural Language Processing (NLP) and expert knowledge elicitation—can scaffold public understanding and participation around complex decisions (Marmolejo-Ramos et al. 2022). At the same time, recent empirical work reports mixed results on whether data storytelling elements consistently improve memory, empathy, or insight detection compared to conventional visualisations, suggesting that narrative features must be carefully designed and empirically tested rather than assumed to be uniformly beneficial (Shao et al. 2024).

Yet, neither visualisation nor narrative, when used in isolation, is sufficient to ensure meaningful understanding. As visualisation research has evolved, interactivity has emerged as a central dimension of public-sector data communication. Advancing traditional approaches that framed interactivity primarily as a tool for analytical exploration, current works reframe it as a site of shared agency, where interpretive authority is negotiated between authors and audiences (Dimara and Stasko 2022; Morini et al. 2025). From a deliberative-democracy perspective, this implies a relevant shift: author-driven narratives orient users toward salient aspects of a policy issue, but exploratory features empower them to interrogate data, challenge institutional framings, and bring situated experiences into the interpretive process. Critical interactivity—bridging narration and exploration—thus enables multiple ways of knowing and aligns with democratic ideals of pluralism, contestation, and epistemic inclusion.

Advancing in the discourse, ethically robust data narratives require transparency about data sources (Chap. "Data Ethics"), methodological choices, and limitations, alongside attention to equity, accessibility, and representation. Nevertheless, historical work in public administration emphasises that transparency is not only about access to information but also about its comprehensibility, contextualisation, and explorability (Magnini et al. 2000). Participatory design approaches—engaging citizens, civil servants, and affected groups in shaping narrative structures or visual framing—can help mitigate risks by incorporating multiple perspectives and knowledge forms (Dove and Jones 2012).

1.4 Data Narratives as Epistemic Infrastructures

Within deliberative democracy, data narratives function as epistemic infrastructures: they shape how issues are problematised, which options seem viable, and how citizens reason together. Effective narrative visualisations can broaden deliberation by making complex evidence accessible, enabling participants to grasp causal mechanisms, evaluate policy trade-offs, and articulate informed arguments. They also strengthen institutional trust by revealing how decisions are grounded in evidence.

Conversely, when narratives present a single interpretive path or closely follow institutional logics, they risk narrowing public debate and excluding experiential or minority knowledge. Critical interactivity offers a counterbalance, allowing users to navigate flexibly between guided stories and open-ended exploration, thus supporting co-interpretation and richer collective sense-making.

Table 1 synthesises the principles emerging from the scientific debate. They operate as foundations for democratic and inclusive public-sector data storytelling and prepare the ground for how AI techniques can further enhance understandability in complex, pluralist, data-rich decision-making environments.

2 The Role of AI in Enhancing Understandability

Given the scenario outlined so far, it is clear that with public-sector decisions increasingly relying on algorithmic systems—from policy analytics to deliberative platforms—the question is no longer only whether AI can analyse, classify, or predict, but whether it can help people understand complex information in ways that are democratically meaningful. In deliberative settings, explainability is inseparable from epistemic accessibility: which publics can understand AI-generated insights, interrogate them, or integrate them into collective reasoning? This raises a critical gap between how Explainable Artificial Intelligence (XAI) (Holzinger 2018; Humer et al. 2024) conceptualises explanation and what democratic contexts actually need.

2.1 The Dominant Paradigm of XAI: Introspection Over Interpretation

As AI becomes increasingly embedded in public-sector analytics, decision-support systems, and civic technologies, the question of understandability extends beyond visualisation to the interpretive layers that machine-generated insights require. A large portion of the research addressing this challenge comes from XAI, which aims at making aspects of an AI system more understandable and interpretable for its intended users. Importantly, though, explanations in this field are conceptualised not as narrative clarifications in the everyday sense but as outputs of a computational system aimed at increasing understandability, appropriateness, and exploitability of AI-generated suggestions and outputs.

Much of the technical XAI literature—summarised in Altukhi et al. (2025) and Kalasampath et al. (2025) and systematised in the manifesto by Longo et al. (2024)—frames indeed explainability as the production of clarifying outputs about model behaviour: feature-importance plots, rule extraction, local counterfactuals, or visualisations of internal model states. The focus is thus on model-centred explanation, presenting it as the main response to the opacity of algorithmic systems—visualising

Table 1 Principles for democratic and understandable data narratives for public-sector uses

Principle	How it is discussed in the literature	Key references	Relevance for democracy & deliberation
Clarity & accessibility	Visualisations must account for heterogeneous levels of data and visualisation literacy. Cognitive load, perceptual constraints, and misinterpretation risks are persistent challenges. Narrative scaffolds and intuitive designs support comprehension and recall	Börner et al. (2016), Maltese et al. (2015), Firat et al. (2022), Shao et al. (2024), Amini et al. (2015)	Ensures that all participants—not only technically skilled individuals—can meaningfully engage with evidence. Reduces epistemic inequality and supports inclusive deliberation by enabling broad comprehension of policy-relevant information
Transparency & reflexivity about data sources and omissions	Data embodies institutional decisions about categories, granularity, and inclusion. Omissions ('dark data') and selective simplifications ('datawash') reproduce power asymmetries and institutional biases if left unexamined	Boehnert (2015, 2016), Magnini et al. (2000)	Makes visible the institutional choices embedded in datasets. Strengthens accountability and public trust by revealing uncertainty, limits, and methodological assumptions critical for democratic scrutiny
Plurality & contestability of interpretations	Narrative visualisations can both guide interpretation and limit it. Research emphasises the need to present multiple perspectives, alternative scenarios, and uncertainty to avoid overly linear or persuasive storytelling. Interactivity helps open interpretive pathways	Tong et al. (2018), Dörk et al. (2013)	Allows citizens to question interpretations, compare perspectives, and surface lived experiences. Supports pluralism and contestation—core conditions of deliberative democracy

(continued)

Table 1 (continued)

Principle	How it is discussed in the literature	Key references	Relevance for democracy & deliberation
Alignment with deliberative processes	Visualisations shape the epistemic environment of public discourse, influencing how issues are framed and what solutions appear viable. Data narratives affect sense-making, argumentation, and reasoning	Boswell et al. (2019), Elstub and Escobar (2019), Vromen (2017)	Ensures that visualisations enable—not constrain—reasoning, dialogue, and collective judgment. Makes complex evidence usable within forums, assemblies, or participatory processes
Shared agency through critical interactivity	Interactivity is understood not only as a technical feature but as a negotiation of authority between designers and users. Critical interactivity bridges author-driven narration and user-driven exploration, allowing audiences to construct their own interpretive pathways	Morini et al. (2025), Dimara and Stasko (2022), Shneiderman (2003)	Empowers citizens to interrogate data, challenge institutional framings, and engage with evidence on their own terms. Supports co-interpretation and richer collective sense-making—central to democratic deliberation

feature importance, surfacing decision rules, or exposing internal model states. These are valuable for debugging and auditing, yet they are designed primarily for experts. This attitude is confirmed by the well-known AI4VIS survey on AI approaches for data visualisation (Wu et al. 2022) which reports state-of-the-art visual analytics pipelines overwhelmingly relying on introspection-oriented techniques that presuppose advanced statistical or ML literacy—thresholds that most citizens and many policymakers cannot reasonably meet.

A core limitation of this paradigm lies in the technical nature of contemporary AI systems themselves. Modern ML models—especially deep neural networks and LLMs—are frequently described as 'black boxes', as their internal representations and decision pathways cannot be directly inspected or intuitively traced (Burrell 2016; von Eschenbach 2021). These architectures operate through high-dimensional, multi-layered transformations that remain opaque even to experts. Importantly, this opacity tends to increase as models become more accurate: state of the art systems often achieve high performance precisely because they exploit patterns and interactions too complex for humans to interpret, reflecting structural tension between predictive accuracy and explainability (Lipton 2018; Rudin 2019). In response, XAI research has developed a wide array of methods to provide insights or explanations,

they fall mainly into two categories: transparency and post-hoc interpretation. Transparent or interpretable-by-design models make their internal logic accessible: their structure, features and decision rules can be examined and understood without auxiliary methods (Murdoch et al. 2019; Rudin 2019). Examples include decision trees, sparse linear models, rule-based systems, or generalized additive models. By contrast post-hoc interpretations are explanatory artefacts generated after an opaque, black-box model has produced an output. Techniques such as saliency maps, counterfactuals, feature attribution, LIME or SHAP do not reveal how the model works; rather they provide approximations or simplified narratives around its behaviour (Doshi-Velez and Kim 2017; Gilpin et al. 2019; Lipton 2018). While post-hoc explanations can be useful for sense-making, they remain exposed to risks of incompleteness, instability, or even contradiction, offering only a partial or potentially misleading picture of the underlying mechanisms (Jacovi et al. 2021).

As a matter of fact, much of the existing work focuses on model-centred transparency—exposing internal parameters, feature importance scores, or decision rules—rather than supporting the broader interpretive, contextual, and deliberative work that enables people to make sense of public problems. Growing empirical evidence also shows that more explanation does not necessarily mean more understanding, nor improve human decision-making: explanations can mislead as much as they clarify (Cabitza et al. 2024). Recent studies are highlighting how users frequently over-trust systems simply because their internal logic appears intelligible and transparent—a dynamic that takes the name of white-box paradox—regardless of whether its explanation is correct or reflects actual model reliability (Bansal et al. 2021). Cabitza et al. (2024) introduce the concept of XAI Halo Effect, where persuasive but flawed explanations degrade human judgment even if the AI advice is correct—as to say that users are influenced by AI-generated misleading explanations to the point that they not verify the correctness of the output. These findings challenge the persistent assumption across XAI research—also widespread in public-sector AI deployment—that transparency automatically improves hybrid human–AI decision-making by fostering better understanding, trust, or decision quality.

Such work points to a crucial conclusion: current forms of explanation typically offered in XAI research often fall short of the needs of democratic governance. Across technical and design literature, the current debate shows indeed a tension between what explainability is assumed to mean in AI research and what understandability requires in democratic contexts. The priority is indeed not understanding the machine, but understanding the issue to which the machine is applied—its uncertainties, trade-offs, and value implications.

2.2 AI as a Facilitator: Narrative, Contextual, and Value-Sensitive Reasoning

Alongside this AI-centred work, a second strand of research—emerging primarily from AI for data visualisation (AI4VIS area) and visible across visual analytics, civic tech, and natural language processing—focuses not on explaining models, but on improving users' ability to understand the data itself. It more specifically looks into how AI can support sensemaking by helping users navigate complex data environments, identify salient patterns, or structure narrative (Wu et al. 2022) show that ML techniques are used to detect trends, rank alternative encodings, or propose data transformations, thereby assisting users in identifying patterns they might otherwise miss. As previously discussed, part of the discussion concerns how AI can structure narrative visualisations by selecting representative frames, sequencing content, and orchestrating transitions (Tong et al. 2018). Adding to this, there is the discourse on how adaptive systems can generate semantic hints, contextual call-outs, and automated highlights that reduce cognitive load and scaffold user interpretation in dashboards (Shao et al. 2024). These functions do not 'explain the model', they provide 'soft guidance' in interactive dashboards, providing adaptive annotations, highlights, or semantic hints that reduce cognitive load and guide user attention. As such they support interpretive work by helping and empowering users navigate complexity with greater clarity, thus contributing to value-sensitive interpretation.

The narrative-building literature complements this view (Bach et al. 2018; Li et al. 2025; Wu et al. 2022). Although still limited, existing examples show how NLP techniques can support public engagement by structuring bottom-up concept clusters, identifying shared values, or revealing latent disagreement—effectively operating as amplifiers of deliberative narrative processes (Anastasiou and De Liddo 2023; Marmolejo-Ramos et al. 2022; Yeo et al. 2024). Techniques such as semantic clustering, sentiment analysis, and expert-knowledge elicitation demonstrate how AI can surface discursive patterns that help publics grasp the broader argumentative landscape. In these cases, AI-generated summaries or clusters can foreground disparities, structural conditions, or areas of uncertainty, making otherwise invisible social patterns more visible for public reasoning.

Nonetheless, these AI-assisted approaches face structural limits. AI models often rely on internal representations designed for machines rather than people, making their outputs difficult for non-experts to interpret (Wu et al. 2022). Deep-learning systems encode information in abstract numerical forms that do not map onto human concepts (Bengio et al. 2013). As a result, AI-generated insights may appear precise yet remain cognitively opaque, limiting their usefulness in democratic decision-making; attempts to visualise these internal representations (Olah et al. 2017) still result in outputs that are unintelligible to most users.

For democratic contexts, this has major implications. If AI systems generate insights based on internal structures that even domain experts struggle to interpret, public-sector users—policymakers, civil servants, and citizens—face unavoidable barriers in assessing the trustworthiness and accountability of results. Moreover,

as Boehnert (2016) observes, model-generated summaries risk reproducing forms of digital positivism: clean, numerical renderings that obscure underlying political, social, or structural determinants. The gap between machine-friendly and human-friendly representations therefore highlights a fundamental challenge: AI is powerful at detecting patterns, but remains limited in supporting the explanatory, contextual, and value-laden reasoning necessary for democratic decision-making.

For these reasons, scholars caution that additional interpretive layers—such as narrative framing, contextual cues, or interactive sensemaking tools—are essential to prevent AI from producing results that appear precise and authoritative while remaining opaque and disconnected from how people actually think, reason, and deliberate about public issues.

Taken together, these strands indicate that AI can meaningfully assist understandability, but only when explanation is conceived as a support for critical engagement—grounded in understandable data, contextual knowledge, and participatory reasoning.

This opens further questions, which add to the ones opening this chapter, which the ORBIS and KT4D cases address in the next sections: How can AI help users navigate complexity without oversteering interpretation? How can explanations reveal diversity of perspective and keep their variety rather than smoothing them over? What capacities do users need to develop in order to be able to become empowered with respect to their technology-mediated interactions? And what forms of interaction preserve agency and contestability in AI-assisted democratic processes?

2.3 Explainability as Democratic Infrastructure: From Situation-Aware Explanations to Participatory Co-assessment

Building on these questions, emerging conceptual challenges must be transformed to concrete design and evaluation responses that have emerged across recent democracy-oriented AI initiatives. If explainability is to function as a support for critical engagement rather than as a mechanism of cognitive steering, it must be grounded in contextual awareness, inclusive evaluation, and governance arrangements that preserve agency and contestability. This requires moving beyond generic notions of explanation toward approaches that recognize the situated nature of understanding and the diversity of stakeholders involved in democratic processes. In this context, explainability should be examined both as a technical property of AI systems and as a democratic infrastructure that enables users to interrogate, contest, and meaningfully engage with algorithmic outputs. This subsection explores how situation-aware explanation strategies, combined with participatory co-assessment practices, can respond to the challenges outlined above by aligning explanation quality with users'

contexts, capacities, and values. In doing so, it illustrates how explainability, inclusiveness, and accountability can be jointly operationalized to support empowerment and pluralism in AI-assisted democratic settings.

For example, one of the key results introduced in the context of the AI4GOV project is the Situation Aware eXplainability (SAX). At its heart is the perception that explainability promotes trust and adoption of automation technology. SAX focuses on tailoring explanations to specific contexts of use, recognizing that the same explanation can be perceived differently depending on situational factors such as the task, stakeholder background, or decision stakes. To efficiently evaluate the quality of explanations generated by a Large Language Model (LLM), the project developed a set of scale metrics based on multiple measurement dimensions. Currently, no universally accepted framework or metrics exist for evaluating perceived explanations. The literature lacks consensus on what constitutes an effective explanation, which properties make explanations understandable, and how these qualities can be systematically measured (Carvalho et al. 2019; Elkhawaga et al. 2023; Markus et al. 2021; Naveed et al. 2024; Sokol and Flach 2020; Vilone and Longo 2021; Zhou et al. 2021). This gap was addressed by focusing on the intrinsic qualities and content of explanations, rather than their hedonic or interaction-based aspects. While some attitude-related factors were considered, these were treated as moderating variables influencing users' perceptions of explanation quality rather than direct determinants. Following standard practices (Kulesza et al. 2013; Markus et al. 2021), the project conducted a user study employing a tailored evaluation scale adapted from existing explainability frameworks. Common dimensions from the literature were first reviewed and then refined to fit this context. Fidelity and interpretability were adopted as two primary latent constructs, each encompassing measurable sub-dimensions. Fidelity included completeness, soundness, and causability, while interpretability comprised clarity, compactness, and comprehensibility. Causability reflects the correctness of the model's internal reasoning and aligns with fidelity, whereas comprehensibility captures the user's ability to understand the explanation, linking to interpretability. Furthermore, interpretability ensures that automated recommendations can be contested and debated, which is an essential feature in democratic processes that value plurality and disagreement as productive forces rather than as errors to be minimized. At the same time, inclusiveness and equity should be preserved to ensure that AI systems support diversity and prevent the reproduction of social inequalities through their design. Algorithms must be developed and trained on representative datasets, avoiding biases that silence or mischaracterize certain social groups. Inclusiveness also extends to participatory design, where different stakeholders, such as citizens, policymakers, experts, journalists, and marginalized communities, are involved in shaping system objectives and evaluation criteria. This participatory inclusiveness transforms AI from a top-down instrument of automation into a collaborative infrastructure for democratic co-creation. Finally, governance and accountability mechanisms are crucial for defining responsibility and oversight when automated systems influence public discourse or policymaking. Clear governance structures should delineate who is accountable for data quality, model design, and the consequences of algorithmic decisions. Ethical

review boards, algorithmic audit trails, and legal safeguards contribute to a governance ecosystem that ensures AI tools remain subordinate to democratic control and human judgment.

Aligned with these values, a major lesson that derived from the abovementioned projects is the role of co-assessment. Unlike typical top-down evaluations, co-assessment should involve policymakers, technologists, data scientists, citizens, and other stakeholders that collaboratively assess system performance and alignment with expectations and fundamental human values and rights. Although the organisation of such an assessment framework will inevitably require more time and effort to organise, observing the operation of a democracy-facing system from multiple perspectives is the only way to deliver a true assessment of their function. The manner in which such a framework might be delivered can be seen in the KT4D programme of Use Cases (see full report on these at Edmond et al. 2024), which followed a matrix approach so as to be able to assemble a variety of perspectives and modes of feedback at different points in the project's development. The four Use Cases were designed to represent three different perspectives on technology design and deployment: regulators and policymakers, citizen end users, and technology designers and builders. Each Use Case also represented a different cultural milieu, with the policy lens being featured in Brussels, the technology development lens in Dublin, and the citizen lens in two cities representing very different linguistic, cultural and political contexts, Madrid and Warsaw. As a final characteristic of this framework, the project's interactions with these communities were designed to grow and develop over time, starting with a participatory design phase, moving through a lab phase, and ending with activities to validate the project's final results. This structure ensured that the project was able to always assess its progress not just from a single user perspective or moment in time, but to triangulate between value-sets in a continuous and agile manner.

3 Building Critical Digital Literacy Through Culture and Agentic Play. Addressing Gaps in Empowerment and Knowledge in KT4D

Democracy is ultimately about people, even when technology plays a role in mediating their interactions. Too often, however, technology users are held responsible for their actions in contexts where they have neither the sense of empowerment nor the depth of knowledge they would require in order to act in an informed manner, fully in line with the best representation of their range of individual and relational values. If they are to be widely adopted, technology tools and platforms should be simple to use, but simplifications in user interface or in the presentation of results can often obscure important design decisions and hamper responsible, holistically self-interested, technology use. From this underlying tension between transparency and simplicity rises a fraught venue for potentially competing interests, namely between

the developers of technology and their users. This friction seems to be quite deeply rooted: already in the 1960s, work by Cannon and Perry indicated that computer programmers '… don't like people—they dislike activities involving close personal interaction; they generally are more interested in things than in people' (Cannon and Perry 1966, p. 63). One might like to believe that the growth of computer science as a discipline and software development as a sector of the economy might have shifted this positioning, but more recent work, such as Birhane et al.'s study of justificatory and value-focussed language in machine learning research (Birhane et al. 2022), indicates that the interest in things over people endures among those who build the digital platforms that frame our cultural participation.

This is not to say that considerations of the ethics of computer science have not evolved, but rather that ethics still have a tendency to be seen as an addition to software platform design, rather than an intrinsic part of it. This may in part be due to a clash in epistemic cultures, with engineering requiring clarity and precision, and ethics often coming into its own only in those grey areas where there is seldom clear right and wrong. Similarly, ethics, having its origins in the traditions of philosophy, tends to enact its findings in description and discussion, while computer science expresses itself more through code and the data processing results that code can deliver. According to the results of the KT4D participatory design session with software designers and design managers (Edmond et al. 2024, pp. 55–68), ethical software development is considered to be very important, but how and at what stage in the development process to introduce it remains open to question (confirming the findings of (confirming the findings of Ayling and Chapman 2022). While many checklists and assessment tools are available, these are generally viewed negatively, as they are often not well known, too generalised, not well-tested, not integrated in software design platforms, and/or not well enough aligned to the complexities of real world situations.

The software developers attending the KT4D-sponsored session were also sensitive to the different forms that ethical software might take, and the different perspectives that could shape ethical behaviour. In particular, while the group did feel they had been given some exposure to issues of intersectional identity, and biases arising from factors such as gender and race, they did not necessarily feel they were prepared to address questions of cultural variation and identities, and issues of the local versus the global. In spite of posing significant challenges, these issues tended to end up being ignored as a result of being deemed too complex to deal with in any systematic manner.

This view into the context in which software is developed raises a number of discrete issues in terms of how tensions between democracy and advanced technologies might be eased. These are particularly relevant to addressing two interrelated barriers—the gap of empowerment and the gap in knowledge—to the creation of democracy-facing technologies that are transparent and meaningful. To bridge these gaps requires not just a shift in product safety regulation, but a more fundamental progression in how we view the critical skills people bring with them into their interactions with technology products and decisions regarding their adoption.

Turning first to the gap in empowerment, and indeed to the stickiest type of ethical problem identified in the KT4D design session, we find that the complexity of culture positions it to serve a dual role of both protector of such forces as identity and community (which in turn may serve as protective of democracy), and of a source of complexity and frustration to those who might want to build systems to enhance civic participation. Neither of these two forces is particularly straightforward to address. In his review of perspectives on the intersection of culture and democracy, Inglehart et al. (1998, p. 80) recognises that 'cultural traditions are remarkably enduring and shape the political and economic behaviour of societies,' but demonstrates a remarkably narrow view of culture as found in many of the works he surveys, some of which oversimplify at the macro level, e.g. by dividing the world up into a small number of religion-driven macro-cultures (e.g. Huntington 1996), or at the micro-level, by considering only certain narrow aspects of culture, such as values related to survival versus self-expression (Ingelhardt's own take). Culture, however, resists such reductions, with values, narratives, practices, languages, beliefs, artistic expression, etc. all remaining staunchly entangled, and individuals drawing often from multiple culturally embedded positions in the formation of their identities. It is these culturally informed positions that ultimately inspire tolerances for trust, community ties, discursive preferences, and other such matters that precondition the individual and the collective toward or against civic and democratic participation, including the all- important layer of trust in governments, regulatory institutions, their representatives and their instruments. When technology seeks to mediate these relationships and interactions in a seamless, frictionless fashion, the protective, lubricating, agency-supporting functions of culture can in fact be damaged, rather than strengthened.

Seen as such, culture can be a force for diversity and fragmentation in a pluralistic society, but also one for empowerment and agency, as it maintains the link between communal decision-making and individual values and sensemaking practices. The received design principles for software interfaces and data handling structures optimise for scale, and as such tend either to hide embedded cultural biases or purport to be culturally neutral, a supposed strength that ultimately manifests as a weakness. Does it matter to your sense of agency in the context of civic participation if an interface is available in your mother tongue, or if you can navigate it only through a politically acceptable vehicular? Does it matter if you happen to belong to a demographic (migrant, disadvantaged, etc.) that is less likely to be comfortable with technological tools in general? Does it matter how the design of any platform uses colors, symbols or layout, or presents issues such as consent or data use (which may be read neutrally or as threatening, depending on your experiences with elites and power structures)? The answer in each of these cases is of course that yes, these things might make a difference, but the culture of software development tends to dismiss these subtleties in the effort to create tools that can be deployed at scale, often by international companies spanning numerous cultural spaces, but maintaining a surprisingly homogenous corporate culture (largely young, largely male, open to risk, etc.).

Is there a mechanism that can allay this tension between the local and global, between culture and scale? It is often the individual user who is expected to be able to make the right choices for themself, in spite of having potentially limited access to understanding of how a platform may have been designed (or function), or indeed to the conditions for its use as they are described in the inevitable lengthy—and seldom read—terms of service the user is required to sign. This situation described the gaps in knowledge that might frame citizen technology interactions, which one might also characterise as gaps in digital literacy. There are many theories and approaches to education for digital literacy, but these tend to focus on the acquisition of applied technology skills, and do not necessarily prepare individuals to think critically and in a specific use context about the impact of their adoption of knowledge technologies. It is not (or is not just) technical literacy that is required, however. There is an equally crowded field of approaches related to how education can build the capabilities needed for democratic societies (such as Riddle and Apple 2019). This noisy space also encompasses the inherent moral hazard that by providing education to citizens, they can then be considered fully competent to manage the risks they face (even when they are being actively manipulated in ways that undermine their ability to use what knowledge they have, or when there is a mismatch between that knowledge and the challenges they face). To mitigate against this risk, the work of the KT4D project focuses on the capabilities of 'critical digital literacy' (CDL), that is the skills needed to assess whether a digital tool is likely to be exploitative, harmful, or out of step with the user's values and culture, and whether it is fit for the purpose of fostering democracy, if that is what it is being presented as capable of doing. This work was modelled conceptually on examples like the Finnish MOOC on AI (elementsofai.com), but keeping more of a focus on allowing individuals to untangle social drivers and business models from technological affordances, and feel empowered to exercise restraint where their values and rights might not align with those enacted or implicit in a given deployment.

In this context, it is useful to draw on the wider traditions of multiliteracies, which argue that literacy should be understood not simply as a set of technical or skills but as a socially situated cultural practice (Buckingham 2007). This body of work highlights how literacy is constructed through social contexts in which texts are produced, circulated and interpreted, and that its meanings vary across cultural settings. This perspective broadens formulations of digital literacy that focus on evaluating or using information, and positions literacy as a meaning-making practice that involves critical reflection. In this sense, being 'literate' involves recognising how media create particular values and power relations, and understanding one's own position within these structures. These ideas support a conceptual grounding for CDL that emphasises cultural context, agency, and critical engagement.

In order to provide a solid and workable definition of 'critical digital literacy', we need to focus separately on the three different concepts making up the definition (critical, digital, and literacy) and ask ourselves what they mean in the context of civic and democratic participation. First, we understand the ability to be 'critical' as necessary for democracy. Indeed, our claim is that people's need to understand and question knowledge technologies, thereby developing a sense of agency over them, is

not only a prerequisite for democratic engagement, but rather indistinguishable from it. Researchers have identified a wide range of approaches to describing this essential component of critical digital literacy, such as, for example, the ability to use media 'to analyse, critique, and transform the norms, rule systems, and practices governing the social fields of institutions and everyday life' (Luke 2014, p. 20). Regarding this notion, there is indeed a general emphasis on gaining self-awareness of one's position in the world, and the way in which it has been historically constructed, and situated within specific power relations (White and Cooper 2015, p. 22). In this way, criticality can be seen as a central strategy by which to foster one of what Bartlett (2018) identifies as the 'six key pillars that make democracy work' and that that technology tends to disrupt, namely the need for active citizens who are alert, independent minded and able to make moral judgements.

The second qualifier 'digital', it is more easily definable, as it refers to the kind of knowledge technologies on which the KT4D project focuses, which are AI and big data. Additionally, we believe that AI and big data need to be understood as more than epiphenomena, but rather as essential to civic participation, because of their status as material tools and semiotic frameworks through which people exercise their democratic power. Finally, 'literacy' in the context of the aims of the KT4D project needs to be understood as implying individual and collective agency (autonomy), empowerment, and identity building, and not simply critical awareness and comprehension skills.

Whether the digital literacy in question concerns media, data, or artificial intelligence, for instance, a critical component lies at the core of each one. This component transcends the level of mere awareness, however, aiming instead to contribute to autonomy as a second and complementary objective of critical literacy. As Buckingham (2003), quoted in Pangrazio (2016, p. 165) maintains that 'the goal of critical literacy is 'not simply critical awareness and understanding, it is critical autonomy' (p. 107). In this approach, critical analysis provides opportunities for 'identity work' (p. 109) in which a variety of social identities can be experimented with.' This highlights the link between literacy and identity which is also at the core of our definition of 'knowledge technologies' expressed with Fig. 1.

As Pangrazio (2016) describes it, CDL can enable a 'dispassionate, critical disposition' and a 'more nuanced understanding of power and ideology within the digital medium.' As such CDL can assist the individual to reconfigure their decision-making with regards to technology use from a mere consumerist consideration of value for

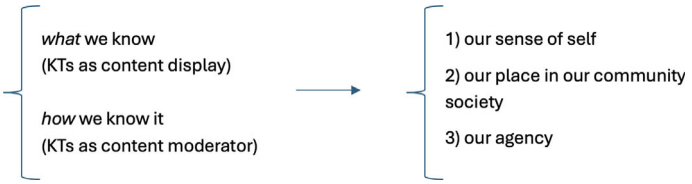

Fig. 1 Knowledge technologies and their impact on identity development

money to an 'examination of the complex interplay of information processing, software dynamics, linguistic processes, and cultural practices that are at work within these digital platforms' (Pangrazio 2016, p. 12). CDL has the capability to render individuals better able to make informed choices about what technologies they adopt, which they reject, and (perhaps most importantly) what information they require to make this decision. Knowing that this is the desired end state for users, and achieving it, are two very different challenges, however.

Toward this second purpose, the KT4D project adopted the paradigm of agentic play (Edmond et al. 2024) as an optimal avenue to the goal of fostering CDL. Games, and play in general, encourage a state of receptivity to new information and active problem solving, a counterbalance to the state of passivity so often encouraged by the frictionless opacity of technology platforms. This engagement might in and of itself be seen as a force for the enhancement of agency, but within KT4D, we remain also mindful of perspectives that position agency as a collective quality (Chia and Ruffino 2022; Harrell and Zhu 2009; Keogh 2018). This is important, given the quintessential nature of democracy as an endeavor position to enhance collective benefit, and drawing from collective experiences of sensemaking. Sensemaking is not just a collective, however, but a narrative practice (Edmond, forthcoming). In spite of the many perspectives positioning the narrative aspects of interactive games as in tension with their ludic nature, we were also able to draw on the concept of 'narrative agency' defined as the 'ability to position ourselves in relation to implicit narratives that steer our actions, self-understandings, and orientation to the future', and to 'practice agential choice over which narratives we use and how we narrate our lives, relationships, and the world around us' (Meretoja 2023, p. 296). The concept of narrative agency was first developed within the field of narrative studies, and later entered the discussion on narrative interactive games because it offers a solution to the 'overly simplified understanding of agency […] as free will of players' (Harrell and Zhu 2009, p. 44).

The concept of narrative agency applied to the design of serious games offered us two things. First, it defines sensemaking as a narrative practice rooted in culture. This is in line with our understanding of culture as the filter through which people adopt, reject, and make sense of technologies, AI included. With our gamified tools we aimed to encourage users to question the cultural values and assumptions that shape their interactions with AI (in case of citizens) and their design of such tools and systems (in case of software developers). The goal is for them to ultimately realise the link between culture, technology, and democratic participation.

Secondly, the concept of narrative agency recognises narrative performance as a polyphonic, collective process (Cunliffe and Coupland 2012) involving contrasting opinions and points of views. This has to do with the fact that both narrative and agency are positional, as they depend on people's cultural values, personal and professional identities, access to knowledge, and more. Through the actioning of this conceptual web, KT4D was able to develop instruments for the fostering of CDL in a number of audiences with diverse information needs. In particular, we focussed on delivering on this vision via three specific avenues toward the enhancement of CDL, targeting three different baseline levels of technological competence

and confidence, and two different decision-making scenarios. For members of the general public with a generally lower baseline level of CDL, we created an open-ended and forensic experience, encouraging empowerment and agency in the face of unfamiliar technologies through a gamified approach. In this escape-room style game, players are presented with political scenarios in which technology plays a role, and encouraged to imagine what impact a given course of action might have. It situates participants within a national leadership scenario, where they navigate a series of ten ethical dilemmas involving advanced knowledge technologies and governance. This immersive environment encourages players to apply CDL concepts while reflecting on the consequences of decision-making for fundamental rights and democracy, fostering civic awareness and a sense of empowerment.

The second set of assets, which are more overtly educational, give a more technologically-aware user the opportunity to learn about, and play with, the affordances of a recommender system and a deepfake video. These more technology-, rather than scenario-led, assets focus directly on the specific challenges of identifying and assessing the role of big data and AI in a knowledge technology by walking users through a set of typical interactions and encouraging consideration at each step. The deepfake explainer takes users through how AI-generated deepfake content is created, how to spot it, and what its broader societal effects might be, encouraging reflection on truth, bias, and the ways our emotions shape how we interpret digital media. The recommendation algorithms explainer lets participants experience a simulated social media feed, showing how their choices influence what content is prioritised or filtered, and providing feedback on what this means in real-world contexts. Seen together, these tools complement the scenario-based escape-room game, with the game focusing on situational and ethical reasoning, while the explainers develop a more procedural understanding of specific technologies and their social implications. By combining immersive, participatory, and reflective experiences, these assets offer a multi-layered approach to Critical Digital Literacy, helping users think critically, make informed decisions, and engage with digital platforms in ways that are both responsible and attuned to social and ethical considerations.

Finally, KT4D also produced an interactive digital narrative aimed at software developers. Here the goal was to develop a specific kind of CDL, namely an awareness of how technology might operate (and do harm) in a specific cultural environment. Our approach is theoretically grounded in what scholars have termed the 'third wave of AI ethics' (Bolte and van Wynsberghe 2025), which transcends technological solutionism to analyse deeper systemic power structures. By adopting a critical digital humanities approach, we challenge developers to perceive their work not merely as code, but as complex cultural interventions with profound societal implications. Central to the methodological approach is a commitment to three key pedagogical and ethical objectives: First, to expose the inherent limitations of algorithmic understanding by creating scenarios that reveal the depth and complexity of cultural communication. Second, to challenge simplistic technological solutions by presenting participants with nuanced, contextually rich interactive experiences that resist reductive interpretations. Third, to provide an immersive learning environment that cultivates critical reflection on the social implications of technological design.

Players are placed in the position of a Red Team that has been hired to determine the cause of a social harm experienced by a client—harm which stems from deep-seated cultural tensions, and which could not be easily attributed to failures by any one party in the system design and delivery. This gamified tool is intended to speak to programmers working in different contexts and at different levels. To choose a specific target—for example, entry level programmers versus senior tech company employees with managerial responsibilities—would have meant not just to narrow the scope of our resource, but especially to overlook the collective and often conflictual negotiations behind AI design.

Each of these assets can be accessed via the KT4D Toolkit.[1] As a suite of materials they support not only the specific CDL needs of citizens and professionals, but point the way toward a revised vision of how digital skills might be understood, and acquired, to the benefit of democratic engagement in the AI age.

4 A Narrative Approach to Data: XAI and Beyond in ORBIS

The ORBIS project provides an empirical demonstration of the theoretical foundations presented in Sects. 1 and 2—concerning literacy, epistemic inequalities, narrative scaffolding, and the limits of traditional XAI—can be translated into an operational approach for democratic deliberation. Rather than attempting to expose or simplify the internal mechanics of LLMs or deep-learning models, ORBIS reframes explainability as a sensemaking challenge, prioritising cognitive accessibility, contextual relevance, and participatory interpretability over technical transparency.

This orientation stems from two mutually reinforcing sources. First, gaps identified in the scientific literature regarding the inadequacy of model-centric XAI in public-sector contexts. Second—and more importantly—the practical needs expressed by ORBIS stakeholders across its project stages (whose co-creation process is detailed in Sect. "Co-creating Innovation in ORBIS"). Scholars highlight the need for more inclusive, participatory, and socially grounded approaches to AI development (Coussement et al. 2024; Siachos and Karacapilidis 2024), emphasising that trustworthy AI requires more than technical safeguards. Indeed, it demands multidisciplinary perspectives, ethical grounding, and meaningful participation throughout the design lifecycle (Delgado et al. 2021; Duberry 2022; Wilson 2022).

ORBIS approach to data understability lies on a robust methodological foundation which combined early co-creation with communities—ranging from citizens to policy practitioners, and civil society organisations—to identify expectations, frustrations, and practical needs in digital deliberation; with iterative co-design with designers, technologists, argumentation scholars, and policy experts, to align on

[1] The toolkit is accessible at: https://kt4democracy.eu/.

how AI should assist, rather than steer, public reasoning. These participatory stages shaped a key guiding principle: explainability is meant for empowering users, not simply interpreting models.

In ORBIS this principle is addressed together with fundamental and well-established insights from critical visualisation scholarship. Dörk et al. (2013) emphasise that visualisation should support disclosure, plurality, contingency, and empowerment, recognising that visualisations are not neutral communicative devices: they mediate meaning through multiple transformations—from world to data, from data to image, and from image to interpretation (Gray et al. 2016). Moreover, interactivity plays a crucial role in how users engage with complex information. It governs the interplay between persons and data interfaces, shaping how users navigate and enact meaning (Dimara and Perin 2020; Dimara and Stasko 2022). Classic principles, such as Shneiderman's (2003) overview first, then zoom and filter, and ultimately provide details on demand, still guide the design of effective visual interfaces, while recent work highlights how interactivity supports narrative visualisation by allowing users to move across story components, explore alternatives, and construct personal interpretive paths (Morini et al. 2025).

4.1 Enhancing Deliberation Data Legibility: Integrating Narrative and Visual Explainability

A fundamental premise requires recalling how deliberative processes generate rich, complex contributions: ideas, claims, counterclaims, emotional reactions, branching argumentation, contextual examples, and personal experiences. When scaled to hundreds or thousands of participants, this data becomes hard to interpret without computational support. However, the challenge is not only scale; it concerns heterogeneity, ambiguity, diversity, and value-ladenness of contributions. Nevertheless, applying algorithmic methods to support data elaboration—topic modelling, clustering, automated summarisation—risks flattening diversity and nuances, up to erasing minority perspectives (Boehnert 2015, 2016).

The conceptual foundations aligned strongly with the needs expressed by ORBIS communities. Ranging across different typologies of participants—youth, activists, policy facilitators, entrepreneurs, and even vulnerable groups—ORBIS early co-creation phase provided clear insights about needs. They don't want technical transparency—namely, how does the model decide?—but rather contextual clarity—how do these data relate? What does this theme mean? How does this argument influence the debate? Participants wanted tools that help them tackle and understand complexity, recognise patterns, and situate their own contributions—not tools that expose the inner mechanics of AI models and LLMs. This directly shaped ORBIS's decision to adopt a Human Centred Explainable AI (HCXAI) (Ehsan et al. 2022; Ehsan and Riedl 2020; Liao and Varshney 2022), reframing explanation as interpretive support rather than model disclosure.

During preliminary workshops, youth participants, civil servants, and community facilitators emphasised requirements which can be summarised as:

1. No need to understand how AI works, but rather how the AI understands them.
2. Need for clear cues about how arguments are classified, grouped, or summarised, particularly when outputs appear unexpected or contestable.
3. Need explanations that are graspable and directly tied to their contributions, not technical descriptions detached from the deliberation's content.

In this setting, and consistently with the conceptual ground established in the previous paragraphs, the ORBIS project confronted a recurring problem in AI-enhanced deliberation: how to make machine-generated insights understandable and democratically meaningful for citizens, facilitators, and policymakers. While Sects. 1 and 2 outlined why visualisation, narrative, context, and explainability matter for deliberative settings, ORBIS offers a concrete example of how these theoretical principles can be translated into practical design decisions for real-world democratic processes.

From the outset, ORBIS faced a dual challenge. On the one hand, large-scale deliberation generates vast, messy, heterogeneous data—free-text arguments, conversational exchanges, emotional reactions, thematic notes. On the other hand, ORBIS AI-enhanced modules—especially the Feedback Aggregator (FA) and the Policy Recommendation (PR), and to some extent some experimentation of Argument Mining (AM)—are powered by LLMs and deep-learning architectures. These models excel at detecting patterns across unstructured discourse, but they produce internal representations that are mathematically coherent for machines yet cognitively opaque for humans. In other words, model-centred explainability does not translate into human-centred understanding.

In response to this gap, ORBIS adopted a human-centred, narrative-driven, participatory approach to explanation. Rather than trying to open the black box of the LLMs themselves, the project reframed explainability as a sensemaking problem. Consequently, ORBIS positioned AI not as a decision-maker, but as a co-analyst that provides structured, explorable visual and narrative aids to understand the deliberative discourse. The aim is not to automate interpretation, but to support participants and moderators in interpreting complex deliberation ecosystems, thus enhancing the identification of insights and patterns in the discourse.

A key component of this reframing is shifting transparency away from the internal logic of the AI models toward the relationship between human contributions and machine-generated structures. Once the goal is no longer to expose how a model computes its internal representations, what becomes democratically meaningful is the ability to see how participants' inputs shape the outputs that organise and render the deliberation intelligible.

In ORBIS, this means foregrounding the interpretive pathway from individual contributions to thematic clusters, summaries, and relational views, allowing users to trace how their arguments, positions, or experiences influence the emerging analytical landscape. By making these correspondences visible and explorable, ORBIS provides a form of transparency anchored in accountability and user agency: participants can

understand not why the model works as it does, but how their own voices and those of others are reflected, contextualised, and assembled into collective patterns.

4.2 From Numerical Indices to Narrative Explanations: Applying Theory to the AM–EG Pipeline

The clearest example of this translation from theory to practice is the iterative design of the Explanation Generator (EG) for the Argument Mining (AM) component.[2] AM analyses each user-submitted statement and assigns it (a) an argumentative type—Position, Support, or Attack, and (b) a functional label—Premise or Claim. The question was how to explain these classifications to participants without reducing them to automation outputs.

1. **The initial approach: Raw feature attribution, unusable for participants**. At the beginning, the AM provided typical XAI-style numerical attributions. specifically, one value per word, ranging from 0 to +1, indicating how much each term contributed to the model's decision. Although this is the canonical approach known in the XAI literature as saliency maps (Lipton 2018), ORBIS testing confirmed what cognitive studies have long shown: numeric introspection is neither informative nor empowering for non-expert users. Participants were confronted with long lists of decimals described as technical noise. No amount of colour-coding improved this interpretability—heatmaps added visual clutter but no semantic clarity. This confirmed insights from human–AI interaction research showing that users rarely interpret raw XAI outputs as designers intend, and may even develop over-trust or misinterpretation (Bansal et al. 2021; Cabitza et al. 2024). This consideration made critical design challenges arise, requiring to explore further the tension between mechanistic transparency and cognitive clarity, particularly regarding the visualisation of the model's confidence values. While the underlying algorithms assign probabilistic weights to distinct textual elements to determine their relevance, visualizing the full spectrum of these values—including low-confidence associations—would arguably enhance transparency by revealing the model's granular decision-making process and inherent uncertainty.

   However, in the context of complex democratic deliberation, such a dense display creates significant visual clutter and extraneous cognitive load, forcing users to decode graphical variables rather than engaging with the core arguments.

[2] The technical development of Explainable Artificial Intelligence (XAI) within ORBIS was led by Elena Cabrio and Sofiane Elguendouze (Université Côte d'Azur) and Serena Villata (National Institute for Research in Digital Science and Technology). However, the strategic decision to move beyond a purely model-centric approach was shaped through the involvement of Ilaria Mariani and Giacomo Garetto (Politecnico di Milano, Department of Design), who contributed expertise on data interpretation and information visualisation. The integration of these design-oriented XAI principles into the platforms was supported by Lucas Anastasiou (The Open University's Knowledge Media Institute).

To mitigate this, the interface employs a strategy of salience thresholding, filtering out low-scoring variables to display only those elements where the model exceeds a definitive confidence level. This design choice deliberately prioritizes pragmatic interpretability over total technical fidelity, increasing the signal-to-noise ratio so that participants can efficiently identify key thematic anchors without being overwhelmed by the underlying probabilistic noise of the system.

2. **The shift to a threshold-based narrative-based model**. The EG has been redesigned in its XAI approach, so that attribution values would still be computed internally, but only the most meaningful terms—those with a significant contribution to the classification—would be shown to users. This thresholding mechanism reduces cognitive load and filters out numerical noise that would overwhelm users. Importantly, thresholds adapt dynamically: for short statements, only one or two key terms may be highlighted; for longer or more complex statements, the explanation may include longer phrase segments or sentence components.
3. **The final narrative layer: from values to meaning**. The filtered attributions are then transformed into a short, human-readable narrative explanation, placed directly under statements like for example:

> The AI model classified this sentence as a Claim based on the contribution of the terms carbon taxes, effective strategies, and reducing greenhouse gas emissions, which directly express a stance on the effectiveness of a policy.

This narrative layer consists of a meaningful solution because it recontextualises machine logic into human concepts, ties explanations directly to user-generated content, supports interpretive agency rather than passive acceptance, and ultimately shifts explainability from model transparency to discursive sensemaking (Miller and Zhang 2024; Yang et al. 2020) (Fig. 2).

4.3 An Integrated Visualisation Design Approach

The redesign of EG so far described was complemented by a suite of visual-narrative tools in BCause, PolisOrbis, and Democratic Reflection. These tools applied the same normative principles—clarity, contextualisation, treaceability—but to higher-level structures of deliberation.

BCause is the platform where this intervention is more substantial and most systematically articulated, since its architecture provides extensive room for advancing and testing visual-narrative approaches. Beyond architecture, its interaction model, workflow, and underlying data structures are specifically designed to support iterative exploration, multi-level sensemaking, and the integration of narrative and visual cues—making BCause particularly suited for implementing and evaluating ORBIS's explainability strategies. This e-participation platform is designed to support one of the core challenges of democratic deliberation: making complex, multi-voiced argumentative landscapes intelligible to participants and facilitators.

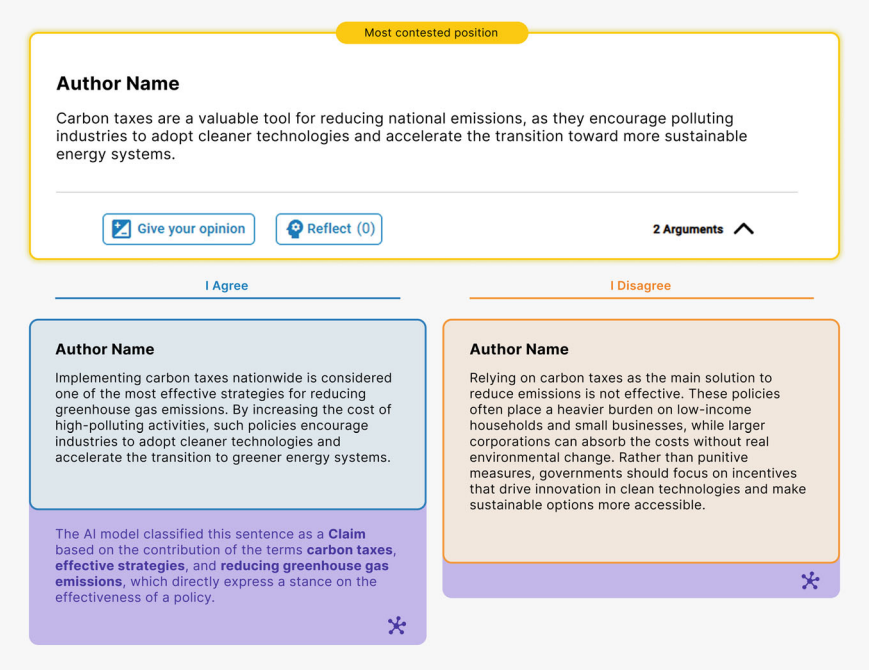

Fig. 2 Concept of explanation generator (EG) applied to the argument mining (AM) component in BCause

Meaningful participation depends on the ability to recognise how individual contributions relate to one another—where positions converge, where disagreements cluster, which arguments support or challenge others, and how themes evolve over time.

BCause addresses this need by transforming dispersed, unstructured inputs into coherent, inspectable structures that reveal patterns of reasoning, areas of consensus and contestation, and the relational dynamics of collective argumentation. Rather than simplifying debate, BCause makes its complexity legible, enabling participants to situate their own voices within the wider discourse and allowing facilitators to guide discussion based on a clearer view of the argumentative setting—ultimately encouraging more informed, reflexive, and inclusive deliberation. To do this, it handles free-text deliberation contributions (opinions, arguments, reactions) in a structured manner and—through the adoption of the ORBIS toolkit—uses AI to elaborate argumentative structures (claims, supports, attacks), thematic clusters, and knowledge-graph visualisations that reveal thematic relations, semantic proximity between contributions, and argument roles within the debate. Through these multi-layered representations, BCause turns patterns that are typically opaque or over-abstracted into inspectable reasoning landscapes, supporting both individual sensemaking and collective facilitation.

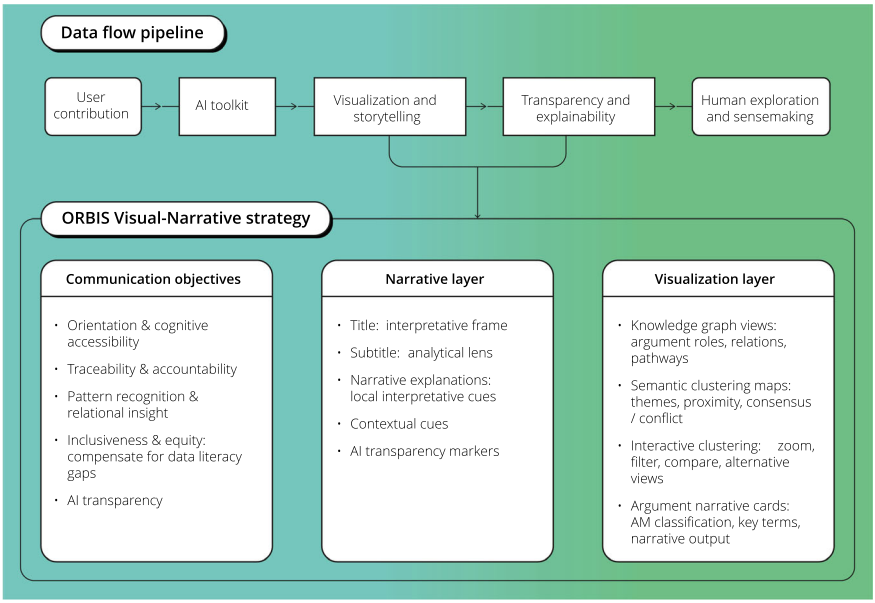

Fig. 3 The multilayered approach for understandability is ORBIS

However, making complex deliberation data visible is not sufficient on its own. As discussed in earlier sections of this chapter, access to knowledge is shaped by uneven levels of data literacy, cognitive load, and representational asymmetries. Without careful design, even the most sophisticated visual analytics can remain inaccessible, overwhelming, or exclusionary. For this reason, the ORBIS team adopted a narrative-driven design strategy in BCause—embedding interpretive scaffolds, contextual cues, and narrative-oriented visualisations to bridge the gap between algorithmic elaboration and human understanding (Fig. 3).

1. **Narrative layer**: Titles and subtitles as sensemaking scaffolds. On the one hand, it centres on enhancing the narrative layer of each visualization through the systematic design of titles and subtitles. These textual elements are conceived as narrative scaffolds and entry points that orient users before they engage with the visual or interactive layer. Research shows that viewers spend the most time on textual elements—especially titles—and that these textual cues strongly shape what users notice, understand, and recall from a visual display (Borkin et al. 2016). In BCause, titles establish the interpretive frame—clarifying what the visualisation is about and what part of the deliberation it represents—while the subtitle introduces the lens through which the data should be read, specifying whether the view presents a synthesis, a relational structure, an area of disagreement, or an exploratory space. This division of roles allowed for encoding both the what (the content surfaced by the AI) and the how (the interpretive stance users should adopt), balancing narrative clarity with explainability. Across

visualisations, this approach ensures consistency while adapting to needs some titles emphasise the human dimension ('Map of the Debate', 'How Arguments Connect'), others make explicit the interpretive function ('Discussion Patterns Exploration'). In parallel, subtitles highlight the contribution of AI ('derived from an analysis of participants' contributions', 'revealing alternative analytical views'). structuring helps users grasp purpose, scope, and epistemic status, reinforcing transparency, reducing ambiguity, and supporting a smoother entry into AI-enhanced data storytelling.

2. **Transparency cues**. To complement narrative clarity, it is introduced a lightweight visual disclaimer whenever a visualisation contains AI-generated content. This cue does not aim to explain the underlying models but ensures that users remain aware that algorithmic processing contributed to the interpretive pipeline. This strengthens reflexivity without burdening users with mechanistic details they neither need nor want.

Moving to the other ORBIS-related platforms, we can observe how the same narrative–visual approach has been adapted and implemented within different deliberative environments, each shaped through co-creation with pilots and transdisciplinary teams.

Democratic Reflection is designed to make live or recorded public debates intelligible through temporal, emotional, and interpretive feedback. During ORBIS this platform was significantly implemented, adopting iterative co-design with facilitators and pilot participants. The implementation ranged from new functionalities and features emerged from pilots, to visual-narrative layer, as relevant per the discourse in this chapter. ORBIS's narrative–visual approach has been introduced in the two dashboards that the platform produces: one for participants and one for facilitators. Participants contributions, generated through reflective flashcards during debates, are analysed to reveal engagement peaks, positions over time, and sentiment patterns; all data visualisations are narrative-anchored, meaning that each point on a chart is clickable and reconnects users to the exact transcript segment from which it originates.

- The **Facilitator Dashboard** provides thematic syntheses, question prompts (open, clarifying, or provocative), timelines of reflective reactions, and speaker-position trajectories along five interpretive dimensions (relevance, agreement, effectiveness, comprehensiveness, sentiment).
- The **User Dashboard** mirrors this structure in an accessible form, offering an annotated transcript, a timeline of reflection cards, summaries of emerging themes, polarity distributions, and interactive links that re-anchor every visual element to its source in the discourse. Through this comprehensive design, Democratic Reflection turns affective and temporal reflections—often the most difficult data to interpret—into coherent, traceable, and participant-centred sensemaking tools.

PolisOrbis extends the original pol.is model—structured around opinion clustering, consensus–divergence detection, and polarity distributions—by integrating the ORBIS Policy Recommendation components of the toolkit and embedding

narrative elements consistent with the project's explainability strategy. Through co-creation with pilot communities and transdisciplinary co-design sessions involving AI specialists, designers, data scientists, and facilitators, PolisOrbis introduced a dedicated interpretive, AI-enhanced report that complements the traditional cluster map provided by Pol.is. Beyond operationalising policy recommendations derived from participants' statements, this report narrates how AI was used and how it contributed to producing actionable insights. In clear, accessible language, it specifies which tools were applied, at what stage, for what purpose, and with what limitations—making the role of AI intelligible without exposing technical internals.

Across all these platforms, the challenge is the same: to transform heterogeneous, large-scale deliberation data into structures that support sensemaking, reflection, comparison, and reasoned judgment. Visualisations therefore become explanatory artefacts, and narrative summaries act as interpretive layers guiding users through complex informational landscapes. ORBIS's approach thus builds directly on the insights of the narrative-visualisation literature (Bach et al. 2018; Shao et al. 2024; Tong et al. 2018) and operationalises them in democratic contexts where cognitive accessibility, equity, and contestability are crucial.

In this light, ORBIS advances the broader discourse on data understandability by demonstrating that integrated approaches to explainability—combining data visualisation, narrative scaffolding, and lightweight interpretive cues—requires to be tailored not only to the technical architecture and features of each platform, but above all to the deliberative processes they are intended to support. This reframing shifts the emphasis from the technology itself to the social needs, reasoning practices, and epistemic conditions of democratic participation, underscoring that explainability becomes meaningful only when it aligns with how people actually deliberate, interpret, and engage with public issues.

This perspective required that the ORBIS toolkit was not merely added to or layered onto existing environments; instead, its integration evolved through careful attention to the specific dynamics of each system, the nature of the data they generate, and the forms of reasoning they seek to cultivate. What emerges across cases is a shared objective: enhancing people's capacity to participate meaningfully in deliberation by enabling them to understand what is being discussed, how contributions relate to one another, and when and how AI is involved in shaping interpretive outcomes. Ensuring such awareness in a graspable, accessible, and fair manner is essential for cultivating trust and legitimate engagement.

At the same time, the ORBIS experience makes clear that there is no one-size-fits-all approach to explainability for democratic contexts. Effective strategies must remain flexible, situated, and responsive to the platform's purpose, the deliberative task at hand, and the needs of diverse stakeholders. This requires designing dedicated narrative–visual solutions that amplify understanding while preserving agency, contestability, and plurality. Ultimately, the lesson that emerges is methodological rather than technical: enhancing deliberation in AI-mediated environments depends on iterative, participatory, context-sensitive design aimed not at producing a perfect explanation, but at generating the most meaningful one for the people involved.

5 Reframing Understandability in Democratic Contexts

Taken together, the KT4D and ORBIS cases offer an empirically grounded response to the theoretical tensions identified at the beginning of this chapter. Much of the current debate in data visualisation, literacy, and explainability acknowledges the structural challenges of epistemic inequality, cognitive overload, and the opacity of algorithmic systems—but empirical work demonstrating how these barriers can be addressed in democratic contexts remains limited. The two EU projects showcased here contribute to this gap, providing evidence of how integrated, socio-technical approaches to understandability can be co-created and co-designed, implemented, and iteratively refined within real deliberative processes.

Both cases challenge the assumption—still dominant in digital literacy frameworks—that citizens' difficulties with AI stem primarily from a lack of technical skills.

KT4D demonstrates that understanding AI in democratic settings is fundamentally cultural, narrative, and agentic. By mobilising concepts such as multiliteracies (Buckingham 2007), narrative agency (Meretoja 2023), and critical digital literacy (Pangrazio 2016), it reframes literacy not as the acquisition of competencies but as the situated ability to question how technologies encode values, identities, and worldviews. The project shows how simulation and agentic play can operationalise this reframing, providing experiential infrastructures through which individuals and groups can test, negotiate, and question the socio-political implications of AI and big data. Aligning with theories of collective and polyphonic agency in interactive systems, literacy becomes a matter of collective sensemaking rather than individual mastery, foregrounding how people reason together about technological power, cultural narratives, and democratic responsibility. This result directly responds to and enriches ongoing scholarly debates on digital empowerment and the politics of technological opacity, by demonstrating how culturally grounded, participatory, and narrative approaches can rebuild user agency in environments shaped by opaque algorithmic systems.

ORBIS extends this contribution on the side of explainability and deliberative data sensemaking. Instead of adopting XAI as the literature largely describes it (Doshi-Velez and Kim 2017; Gilpin et al. 2019; Kalasampath et al. 2025; Lipton 2018), it demonstrates what explainability becomes when it is re-anchored in the epistemic and participatory needs of democratic deliberation. The project confirms that the primary obstacle to comprehension—and to empowerment within deliberative discourse—is not the opacity of the model itself, but the gap between algorithmic representations and the reasoning practices through which people interpret public issues, echoing concerns raised about visual literacy, epistemic inequalities, and representational asymmetries (Boehnert 2015, 2016; Firat et al. 2022; Shao et al. 2024). In this reframing, explanation is the ability for participants to grasp how their contributions shape clusters, summaries, themes, and recommendations, thereby aligning with democratic ideals of transparency, contestability, and epistemic inclusion. ORBIS operationalises this stance through concrete design innovations

co-designed with interdisciplinary teams—engaging facilitators, communities, and policy practitioners. This confirms the claim advanced by scholars such as Dove and Jones (2012), Coussement et al. (2024), and Delgado et al. (2021) that explainability in democratic AI systems cannot be engineered in isolation; rather, it must emerge through participatory, transdisciplinary, and context-sensitive design processes.

In positioning these results against existing literature, a number of advances become clear. First, visualisation becomes explanation: whereas critical visualisation scholarship often calls for richer, value-aware depictions of data, ORBIS shows how visual forms can themselves operate as explanatory devices that reveal how deliberative contributions relate, evolve, and produce meaning. Second, narrative framing becomes interpretation: moving beyond data storytelling as a communicative layer. ORBIS demonstrates that narrative scaffolds actively shape the epistemic accessibility and legitimacy of AI-supported insights. Third, co-creation becomes a method for democratic alignment: responding to critiques about the absence of participatory design in civic-AI systems. Both KT4D and ORBIS show how involving affected publics and practitioners is not merely desirable, but structurally necessary for producing explanations that resonate with lived reasoning practices. Finally, interactivity becomes a mechanism for agency: not simply a navigational aid, but a democratic feature that allows users to challenge interpretations, access underlying discourse, and construct alternative understandings.

These contributions point toward an overarching shift in how understandability should be conceived in democratic contexts. Explainability becomes meaningful only when it supports how people actually deliberate—not how machines compute; literacy becomes empowering only when it enhances collective sensemaking and the ability to participate; and visualisation becomes inclusive and fair only when it creates pathways for interpretation, reflection, and recognition of diverse experiences.

In this sense, the projects discussed here are significant because they articulate a methodological orientation: iterative, participatory, and context-sensitive design as the foundation for any attempt to make AI-mediated democratic processes and data elaboration intelligible and trustworthy.

Ultimately, the chapter points out how there is no single recipe or universal solution for explainability or literacy in democratic settings. Different deliberative processes, institutional contexts, and publics bring distinct needs—and therefore require differentiated approaches, design responses, and socio-technical solutions. What matters is the alignment between these design choices and the epistemic practices they aim to support, achieved through rigorous critical analysis, evidence- and need-based knowledge, and the meaningful engagement of relevant stakeholders. The contribution of KT4D and ORBIS lies precisely in proving that such alignment is not only theoretically necessary but practically achievable. Together, they provide concrete, situated models of how data understandability can be cultivated as a relational capacity—one that emerges between people, technologies, and the public issues they collectively care about.

Fundings The reasoning presented in this work derive from knowledge and insights from four European projects that have received funding from the European Union's Horizon Europe Programme:

1. 'ORBIS. Augmenting participation, co-creation, trust and transparency in Deliberative Democracy at all scales' has received funding under Grant Agreement No. 101094765.
2. 'KT4D. Knowledge Technologies for Democracy' has received funding under Grant Agreement No. 101094302.
3. 'AI4GOV. Trusted AI for Transparent Public Governance fostering Democratic Values' has received funding under Grant Agreement No. 101094905.
4. 'ITHACA. artificial Intelligence To enHAnce Civic pArticipation' has received funding under Grant Agreement No. 101094364.

The opinions expressed herewith are solely of the authors and do not necessarily reflect the point of view of any EU institution.

AI declaration AI tools were employed in the preparation of this chapter solely for proofreading, language refinement, and accessibility enhancement, not for generating original content or research results. The use of AI-based proofreading support (LLM specifically) was particularly valuable given that not all the authors are non-native English speakers, ensuring improved linguistic clarity, consistency, and readability without altering the scientific meaning or interpretation of the content.

References

Altukhi ZM, Pradhan S, Aljohani N (2025) A systematic literature review of the latest advancements in XAI. Technologies 13(3). https://doi.org/10.3390/technologies13030093

Amini F, Henry Riche N, Lee B, Hurter C, Irani P (2015) Understanding data videos: looking at narrative visualization through the cinematography lens. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp 1459–1468. https://doi.org/10.1145/2702123.2702431

Anastasiou L, De Liddo A (2023) BCause: reducing group bias and promoting cohesive discussion in online deliberation processes through a simple and engaging online deliberation tool. In: Chawla K, Shi W (eds) Proceedings of the first workshop on social influence in conversations (SICon 2023). Association for Computational Linguistics, pp 39–49. https://doi.org/10.18653/v1/2023.sicon-1.5

Ayling J, Chapman A (2022) Putting AI ethics to work: are the tools fit for purpose? AI Ethics 2(3):405–429. https://doi.org/10.1007/s43681-021-00084-x

Bach B, Stefaner M, Boy J, Drucker S, Bartram L, Wood J, Ciuccarelli P, Engelhardt Y, Koeppen U, Tversky B (2018) Narrative design patterns for data-driven storytelling. In: Data-driven storytelling. AK Peters/CRC Press, pp 107–133. https://doi.org/10.1201/9781315281575-5

Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI conference on human factors in computing systems. https://doi.org/10.1145/3411764.3445717

Bartlett J (2018) The people versus tech: how the internet is killing democracy (and how we save it). Penguin Random House

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Birhane A, Kalluri P, Card D, Agnew W, Dotan R, Bao M (2022) The values encoded in machine learning research. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, pp 173–184. https://doi.org/10.1145/3531146.3533083

Boehnert J (2015) The politics of data visualisation. Discov Soc 23. https://westminsterresearch.westminster.ac.uk/item/9w5yx/the-politics-of-data-visualisation
Boehnert J (2016) Data visualisation does political things. In: DRS2016: design+ research+ society: future-focused thinking. https://drs2016.squarespace.com/387
Bolte L, van Wynsberghe A (2025) Sustainable AI and the third wave of AI ethics: a structural turn. AI Ethics 5(2):1733–1742. https://doi.org/10.1007/s43681-024-00522-6
Borkin MA, Bylinskii Z, Kim NW, Bainbridge CM, Yeh CS, Borkin D, Pfister H, Oliva A (2016) Beyond memorability: visualization recognition and recall. IEEE Trans Visual Comput Graphics 22(1):519–528. https://doi.org/10.1109/TVCG.2015.2467732
Börner K, Maltese A, Balliet RN, Heimlich J (2016) Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. Inf vis 15(3):198–213. https://doi.org/10.1177/1473871615594652
Boswell J (2013) Why and how narrative matters in deliberative systems. Polit Stud 61(3):620–636. https://doi.org/10.1111/j.1467-9248.2012.00987.x
Boy J, Rensink RA, Bertini E, Fekete J-D (2014) A principled way of assessing visualization literacy. IEEE Trans Vis Comput Graph 20(12):1963–1972. https://doi.org/10.1109/TVCG.2014.2346984
Bruner J (1991) The narrative construction of reality. Crit Inq 18(1):1–21. https://doi.org/10.1086/448619
Buckingham D (2003) Media education: literacy, learning and contemporary culture. Polity Press
Buckingham D (2007) Digital media literacies: rethinking media education in the age of the internet. Res Comp Int Educ 2(1):43–55. https://doi.org/10.2304/rcie.2007.2.1.43
Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 3(1):2053951715622512. https://doi.org/10.1177/2053951715622512
Cabitza F, Fregosi C, Campagner A, Natali C (2024) Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-AI decision making. In: Longo L, Lapuschkin S, Seifert C (eds) Explainable artificial intelligence. Springer Nature Switzerland, pp 255–269. https://doi.org/10.1007/978-3-031-63803-9_14
Cannon WM, Perry DK (1966) A vocational interest scale for computer programmers. In: Proceedings of the fourth SIGCPR conference on computer personnel research, pp 61–82. https://doi.org/10.1145/1142620.1142628
Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8):832. https://doi.org/10.3390/electronics8080832
Chia A, Ruffino P (2022) Special Issue Introduction: politicizing agency in digital play after humanism. Convergence 28(2):309–319. https://doi.org/10.1177/13548565221100135
Claes S, Vande Moere A (2017) The impact of a narrative design strategy for information visualization on a public display. In: Proceedings of the 2017 conference on designing interactive systems, pp 833–838. https://doi.org/10.1145/3064663.3064684
Coussement K, Abedin MZ, Kraus M, Maldonado S, Topuz K (2024) Explainable AI for enhanced decision-making. Decis Support Syst 184:114276. https://doi.org/10.1016/j.dss.2024.114276
Cunliffe A, Coupland C (2012) From hero to villain to hero: making experience sensible through embodied narrative sensemaking. Hum Relat 65(1):63–88. https://doi.org/10.1177/0018726711424321
Delgado F, Yang S, Madaio M, Yang Q (2021) Stakeholder participation in AI: beyond 'add diverse stakeholders and stir'. https://arxiv.org/abs/2111.01122
Dimara E, Perin C (2020) What is interaction for data visualization? IEEE Trans Visual Comput Graph 26(1):119–129. https://doi.org/10.1109/TVCG.2019.2934283
Dimara E, Stasko J (2022) A critical reflection on visualization research: where do decision making tasks hide? IEEE Trans Visual Comput Graph 28(1):1128–1138. https://doi.org/10.1109/TVCG.2021.3114813
Dörk M, Feng P, Collins C, Carpendale S (2013) Critical InfoVis: exploring the politics of visualization. In: CHI '13 extended abstracts on human factors in computing systems, pp 2189–2198. https://doi.org/10.1145/2468356.2468739

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608
Dove G, Jones S (2012) Narrative visualization: sharing insights into complex data. In: Interfaces and human computer interaction (IHCI 2012). https://openaccess.city.ac.uk/id/eprint/1134/
Duberry J (2022) AI and civic tech: engaging citizens in decision-making processes but not without risks. Edward Elgar Publishing, pp 195–224. https://doi.org/10.4337/9781788977319.00012
Echeverria V, Martinez-Maldonado R, Buckingham Shum S (2017) Towards data storytelling to support teaching and learning. In: Proceedings of the 29th Australian conference on computer-human interaction, pp 347–351. https://doi.org/10.1145/3152771.3156134
Edmond J (2026) From 'Hello world' to 'How are we doing?': fostering collective sensemaking with and around the deafness of code. AI Soc. https://doi.org/10.1007/s00146-026-02991-1
Edmond J, Lima E, Martínez CG (2024) Fostering agentic play between technology and democracy. In: 2024 26th international symposium on symbolic and numeric algorithms for scientific computing (SYNASC), pp 304–309. https://doi.org/10.1109/SYNASC65383.2024.00057
Ehsan U, Riedl MO (2020) Human-centered explainable AI: towards a reflective sociotechnical approach. In: Stephanidis C, Kurosu M, Degen H, Reinerman-Jones L (eds) HCI International 2020—Late breaking papers: multimodality and intelligence. Springer International Publishing, pp 449–466. https://doi.org/10.1007/978-3-030-60117-1_33
Ehsan U, Wintersberger P, Liao QV, Watkins EA, Manger C, Daumé III H, Riener A, Riedl MO (2022) Human-centered explainable AI (HCXAI): beyond opening the black-box of AI. In: Extended abstracts of the 2022 CHI conference on human factors in computing systems. https://doi.org/10.1145/3491101.3503727
Elkhawaga G, Elzeki O, Abuelkheir M, Reichert M (2023) Evaluating explainable artificial intelligence methods based on feature elimination: a functionality-grounded approach. Electronics 12(7):1670. https://doi.org/10.3390/electronics12071670
Elstub S, Escobar O (2019) Defining and typologising democratic innovations. In: Elstub S, Escobar O (eds) Handbook of democratic innovation and governance. Elgar, pp 11–31. https://doi.org/10.4337/9781786433862.00009
Fareed N, Swoboda CM, Chen S, Potter E, Wu DT, Sieck CJ (2021) U.S. COVID-19 state government public dashboards: an expert review. Appl Clin Inform 12(02):208–221. https://doi.org/10.1055/s-0041-1723989
Firat EE, Joshi A, Laramee RS (2022) Interactive visualization literacy: the state-of-the-art. Inf vis 21(3):285–310. https://doi.org/10.1177/14738716221081831
Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2019) Explaining explanations: an overview of interpretability of machine learning. https://arxiv.org/abs/1806.00069
Goñi J (2025) Citizen participation and technology: lessons from the fields of deliberative democracy and science and technology studies. Human Soc Sci Commun 12(1):287. https://doi.org/10.1057/s41599-025-04606-4
Gray J, Bounegru L, Milan S, Ciuccarelli P (2016) Ways of seeing data: toward a critical literacy for data visualizations as research objects and research devices. In: Kubitschko S, Kaun A (eds) Innovative methods in media and communication research. Springer International Publishing, pp 227–251. https://doi.org/10.1007/978-3-319-40700-5_12
Harrell DF, Zhu J (2009) Agency play: dimensions of agency for interactive narrative design. In: AAAI spring symposium: intelligent narrative technologies II, pp 44–52
Holzinger A (2018) Explainable AI (ex-AI). Informatik-Spektrum 41(2):138–143. https://doi.org/10.1007/s00287-018-1102-5
Hullman J, Diakopoulos N (2011) Visualization rhetoric: framing effects in narrative visualization. IEEE Trans Vis Comput Graph 17(12):2231–2240. https://doi.org/10.1109/TVCG.2011.255
Humer C, Hinterreiter A, Leichtmann B, Mara M, Streit M (2024) Reassuring, misleading, debunking: comparing effects of XAI methods on human decisions. ACM Trans Interact Intell Syst 14(3). https://doi.org/10.1145/3665647
Huntington SP (1996) Democracy for the long haul. J Democr 7(2):3–13. https://doi.org/10.1353/jod.1996.0028

Inglehart RF, Basanez M, Moreno A (1998) Human values and beliefs: a cross-cultural sourcebook. University of Michigan Press

Jacovi A, Marasović A, Miller T, Goldberg Y (2021) Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 624–635. https://doi.org/10.1145/3442188.3445923

Kalasampath K, Spoorthi KN, Sajeev S, Kuppa SS, Ajay K, Maruthamuthu A (2025) A literature review on applications of explainable artificial intelligence (XAI). IEEE Access 13:41111–41140. https://doi.org/10.1109/ACCESS.2025.3546681

Keogh B (2018) A play of bodies: how we perceive videogames. MIT Press

Kulesza T, Stumpf S, Burnett M, Yang S, Kwan I, Wong W-K (2013) Too much, too little, or just right? Ways explanations impact end users' mental models. In: 2013 IEEE symposium on visual languages and human centric computing, pp 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

Li H, Wang Y, Liao QV, Qu H (2025) Why is AI not a panacea for data workers? An interview study on human-AI collaboration in data storytelling. IEEE Trans Visual Comput Graph 31(10):7598–7613. https://doi.org/10.1109/TVCG.2025.3552017

Liao QV, Varshney KR (2022) Human-centered explainable AI (XAI): from algorithms to user experiences. https://arxiv.org/abs/2110.10790

Lipton ZC (2018) The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57. http://doi.org/10.1145/3236386.3241340

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser JD, Guidotti R, Hayashi Y, Herrera F, Holzinger A, Jiang R, Khosravi H, Lecue F, Malgieri G, Páez A, Samek W, Schneider J, Speith T, Stumpf S (2024) Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Inf Fusion 106:102301. https://doi.org/10.1016/j.inffus.2024.102301

Luke A (2014) Defining critical literacy. In: Pandya J, Ávila J (eds) Moving critical literacies forward. Routledge, pp 19–31

Magnini B, Not E, Stock O, Strapparava C (2000) Natural language processing for transparent communication between public administration and citizens. Artif Intell Law 8(1):1–34. https://doi.org/10.1023/A:1008394902165

Maltese AV, Harsh JA, Svetina D (2015) Data visualization literacy: investigating data interpretation along the novice-expert continuum. J Coll Sci Teach 45(1):84–90. https://doi.org/10.2505/4/jcst15_045_01_84

Markus AF, Kors JA, Rijnbeek PR (2021) The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 113:103655. https://doi.org/10.1016/j.jbi.2020.103655

Marmolejo-Ramos F, Workman T, Walker C, Lenihan D, Moulds S, Correa JC, Hanea AM, Sonna B (2022) AI-powered narrative building for facilitating public participation and engagement. Discov Artif Intell 2(1):7. https://doi.org/10.1007/s44163-022-00023-7

Meretoja H (2023) Implicit narratives and narrative agency. In: Narrative inquiry, vol 33, no 2. John Benjamins, pp 288–316. https://doi.org/10.1075/ni.21076.mer

Miller T, Zhang J (2024) Explanation in artificial intelligence: insights from the social sciences. Digit Human Res 4(2):90–128. https://doi.org/10.1007/978-3-031-42682-7_23

Morini F, Garretón M, Pomerance J, Zeissig N, de Guenther S, Thomet F, Freyberg L, Kyriazis I, Scholz A, Dörk M (2025) Critical interactivity: exploration and narration in data visualization. IEEE Comput Graph Appl 45(3):58–72. https://doi.org/10.1109/MCG.2025.3544684

Mörth E, Bruckner S, Smit NN (2023) ScrollyVis: interactive visual authoring of guided dynamic narratives for scientific scrollytelling. IEEE Trans Vis Comput Graph 29(12):5165–5177. https://doi.org/10.1109/TVCG.2022.3205769

Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci 116(44):22071–22080. https://doi.org/10.1073/pnas.1900654116

Naveed S, Stevens G, Robin-Kern D (2024) An overview of the empirical evaluation of explainable AI (XAI): a comprehensive guideline for user-centered evaluation in XAI. Appl Sci 14(23):11288. https://doi.org/10.3390/app142311288

Pangrazio L (2016) Reconceptualising critical digital literacy. Discour: Stud Cult Polit Educ 37(2):163–174. https://doi.org/10.1080/01596306.2014.942836

Pozdniakov S, Martinez-Maldonado R, Tsai Y-S, Srivastava N, Liu Y, Gasevic D (2023) Single or multi-page learning analytics dashboards? Relationships between teachers' cognitive load and visualisation literacy. In: Viberg O, Jivet I, Muñoz-Merino PJ, Perifanou M, Papathoma T (eds) Responsive and sustainable educational futures. Springer Nature Switzerland, pp 339–355. https://doi.org/10.1007/978-3-031-42682-7_23

Riddle S, Apple MW (2019). Re-imagining education for democracy. Routledge London. https://doi.org/10.4324/9780429242748

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. https://arxiv.org/abs/1811.10154

Schneiders P (2020) What remains in mind? Effectiveness and efficiency of explainers at conveying information. Media Commun 8(1):218–231. https://doi.org/10.17645/mac.v8i1.2507

Segel E, Heer J (2010) Narrative visualization: telling stories with data. IEEE Trans Vis Comput Graph 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 systems. https://doi.org/10.1145/3613904.3643022

Shneiderman B (2003) The eyes have it: a task by data type taxonomy for information visualizations. In: Bederson BB, Shneiderman B (eds) The craft of information visualization. Morgan Kaufmann, pp 364–371. https://doi.org/10.1016/B978-155860915-0/50046-9

Siachos I, Karacapilidis N (2024) Explainable artificial intelligence methods to enhance transparency and trust in digital deliberation settings. Future Internet 16(7). https://doi.org/10.3390/fi16070241

Sokol K, Flach P (2020) Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 56–67. https://doi.org/10.1145/3351095.3372870

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

Sweller J, van Merrienboer JJG, Paas FGWC (1998) Cognitive architecture and instructional design. Educ Psychol Rev 10(3):251–296. https://doi.org/10.1023/A:1022193728205

Tong C, Roberts R, Borgo R, Walton S, Laramee RS, Wegba K, Lu A, Wang Y, Qu H, Luo Q, Ma X (2018) Storytelling and visualization: an extended survey. Information 9(3). https://doi.org/10.3390/info9030065

Tufte ER, Graves-Morris PR (1983) The visual display of quantitative information, vol 2. Graphics Press

Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion 76:89–106. https://doi.org/10.1016/j.inffus.2021.05.009

von Eschenbach WJ (2021) Transparency and the black box problem: why we do not trust AI. Philos Technol 34(4):1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Vromen A (2017) Digital citizenship and political engagement. In: Vromen A (ed) Digital citizenship and political engagement: the challenge from online campaigning and advocacy organisations. Palgrave Macmillan, UK, pp 9–49. https://doi.org/10.1057/978-1-137-48865-7_2

White RE, Cooper K (2015) What is critical literacy? In: White RE, Cooper K (eds) Democracy and its discontents: critical literacy across global contexts. SensePublishers, pp 21–35. https://doi.org/10.1007/978-94-6300-106-9_2

Wilson C (2022) Public engagement and AI: a values analysis of national strategies. Gov Inf Q 39(1):101652. https://doi.org/10.1016/j.giq.2021.101652

Wu A, Wang Y, Shu X, Moritz D, Cui W, Zhang H, Zhang D, Qu H (2022) AI4VIS: survey on artificial intelligence approaches for data visualization. IEEE Trans Vis Comput Graph 28(12):5049–5070. https://doi.org/10.1109/TVCG.2021.3099002

Yang Q, Steinfeld A, Rosé C, Zimmerman J (2020) Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp 1–13. https://doi.org/10.1145/3313831.3376301

Yeo S, Lim G, Gao J, Zhang W, Perrault ST (2024) Help me reflect: leveraging self-reflection interface nudges to enhance deliberativeness on online deliberation platforms. In: Proceedings of the 2024 CHI conference on human factors in computing systems. https://doi.org/10.1145/3613904.3642530

Zhang Y, Reynolds M, Lugmayr A, Damjanov K, Hassan GM (2022) A visual data storytelling framework. Informatics 9(4). https://doi.org/10.3390/informatics9040073

Zhang Y, Sun Y, Gaggiano JD, Kumar N, Andris C, Parker AG (2023) Visualization design practices in a crisis: behind the scenes with COVID-19 dashboard creators. IEEE Trans Vis Comput Graph 29(1):1037–1047. https://doi.org/10.1109/TVCG.2022.3209493

Zhou J, Gandomi AH, Chen F, Holzinger A (2021) Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5):593. https://doi.org/10.3390/electronics10050593