Inclusivity in AI-Enhanced Democracy

Jennifer Edmond, Ilaria Mariani, Sabrina Sacco, Francesca Rizzo, Grazia Concilio, Atte Ojanen, Anna Bjork, George Manias, Xanthi Papageorgiou, Nikos Kalantzis, and Konstantinos Tzelaptsis

Abstract This chapter will approach this challenge of balancing the opacity and value of efficiency that often sits at the heart of technology design, with the openness and individual voice required for functional democracy. It will do so by looking at the challenge from a number of different perspectives. The first of these will be the perspective of **technological design** and **data**. AI and big data systems are

J. Edmond (✉)
Trinity College Dublin, Dublin, Ireland
e-mail: edmondj@tcd.ie

I. Mariani · S. Sacco · F. Rizzo
Department of Design, Politecnico di Milano, Milan, Italy
e-mail: ilaria1.mariani@polimi.it

S. Sacco
e-mail: sabrina.sacco@polimi.it

F. Rizzo
e-mail: francesca.rizzo@polimi.it

G. Concilio
Department of Architecture and Urban Studies, Politecnico di Milano, Milan, Italy
e-mail: grazia.concilio@polimi.it

A. Ojanen · A. Bjork
Demos Helsinki, Helsinki, Finland
e-mail: atte.ojanen@demoshelsinki.fi

A. Bjork
e-mail: anna.bjork@demoshelsinki.fi

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

X. Papageorgiou · N. Kalantzis · K. Tzelaptsis
UBITECH Ltd., Limassol, Cyprus
e-mail: xpapageorgiou@ubitech.eu

N. Kalantzis
e-mail: nkalantzis@ubitech.eu

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_6

notoriously closed and opaque, and corporate approaches to the collection of training and testing data have not always respected human values and democratic principles. The second perspective will focus on **policies and principles**, and the way in which we can create a common understanding not just of how these systems should be designed, but the constraints we need to place on how they are used, and with and for whom. Finally, we will look at the way in which both users and policymakers can be encouraged to develop a refined and enhanced sense of **critical digital literacy** so as to improve both the agency of the end user and the policymaker, fostering the emergence of the model of democracy-in-the-loop through the introduction of production friction.

Keywords Inclusivity · AI-enhanced democracy · Critical digital literacy · Evaluation · Assessment · Marginalisation · Regulation · Inclusive democracy · e-Participation

1 What is Inclusivity?

It seems to go without saying that democracy, and any platforms that are designed and built to support it, must be inclusive. Democracy that does not grant voice to all of its citizens is not able to function optimally, as the perception of exclusions and inequality will naturally erode civic participation over time. It is assumed that digital platforms can play a role in helping to manage this risk, ensuring that citizens can each feel heard and that a more productive form of consensus can be achieved through the digital collection and processing of user perspectives, but this can only succeed if those platforms are themselves truly inclusive, and respectful of the human characteristics that are the soul of democratic exchange.

Inclusivity is not a simple matter of shifting technical design, however. Inclusivity is a shifting challenge that requires agility in both AI and big data systems facing democracy and civic participation, but also facing the processes by which these systems are imagined, commissioned and deployed. Only by creating systems that are responsive and transparent, showing not just their inputs, but also the metabolism by which they are transformed into recommendations, including any sources of bias that the systems may have built into them, can platforms support the mission of enhancing, rather than constraining, the agency of citizens to participate in the messiness of democratic exchange.

This chapter will approach the challenge of balancing the opacity and value of efficiency that often sits at the heart of technology design, with the openness and individual voice required for functional democracy. It will do so by looking at the challenge from a number of different perspectives. It will open with a consideration

K. Tzelaptsis
e-mail: ktzelaptsis@ubitech.eu

of the state of inclusivity in the theory and practice of democratic technology development, opening out from there into a discussion of how we might reappropriate the concept of control as a way to represent and create balance in the system of stakeholders that frame the design and deployment of such systems. The chapter will then turn to a series of reflections on specific areas that inclusive technology needs to be mindful of: the safeguarding of human rights and the challenge of identifying and combatting marginalisation.

2 When Technology Enables Inclusive Democracy

According to Overton, an inclusive democracy is ‘a pluralistic, liberal democracy with a racially and ethnically diverse population that protects the political rights of individuals of all groups equally’ (Overton 2025, p. 13). Although he makes his arguments specifically in the context of racial diversity in a US context, his perspective has wider applicability for Europe as well, capturing as it does the values of respect for cultural diversity, recognition of the benefits of interdependence within this diverse context, and the need for equity of voice among social and demographic groupings. How this desired state of inclusivity might be maximised in a technologised age remains a contested question, however. The European Commission’s 2020 Democracy Action Plan presented the protection of election integrity, promotion of democratic participation, strengthening of the media and countering disinformation as the primary challenges to a healthy, functioning democratic environment in the EU bloc. Even before then, however, some presented visions that framed this focus as too narrow. Hofmann (2019), for example, suggested that we cannot disconnect technology from democracy, as the relationship between them is an inherently co-evolutionary one, where democracy should be seen as mediated through digital technologies, on the one hand, and those same technologies also understood as being influenced and shaped by democratic practices and theories on the other. While most of the literature from that period focuses on the impact of AI and big data on freedom of expression, data protection, privacy and other democratic principles, some work went deeper into this relationship of co-creation, examining facets such as the impact of AI on culture and identity-based concepts, including the right to assembly and association (Ashraf 2020a), and the function of an inclusive public sphere (Kruse et al. 2018), which requires that: (1) access to information be limitless; (2) participation be fair, unrestricted, and protected; and (3) influence of institutions, particularly in terms of economics, be controlled. AI and big data have a two-fold impact on the construction of this kind of emerging public sphere, shaping both content access by influencing what we see, and content moderation, whereby AI influences what themes emerge within democratic discussion and deliberation.

This background represents the context for considering the interactions between AI, big data and the requirements of an inclusive democracy from the time when the projects of the AI, Big Data and Democracy Forum were being conceived. Only a few years later, as these projects are releasing their results, the landscape seems very different, with new policy positions, new threats, and radically new technological affordances. The European Commission's 2025 Democracy Shield initiative is a reflection of these shifts, While its headline actions maintain its precursor's focus on the policy and institutional imperatives available to protect democracy, at a more granular level, a perception of technological threat—via disinformation, election tampering, and citizen disempowerment—clearly emerges, and many of the levers indicated, such as the Digital Markets, Digital Services and AI Acts, operate on this level.

The intensity of technology introduction in this period, with the rapid adoption of generative AI tools capable of producing text, images, code, presentations and other artifacts of human-like quality, has been met as well with a rise in technology research and innovation specific to democracy-facing contexts. This body of work, while important, seems at times to struggle with the rapid pace of change, with the perspective of true inclusivity at times seeming to fade from central view. Platforms for democratic exchange have been designed and released, such as BCause, Remesh, po.lis, Decidim, deliberation.io, and many others, but not all of the entries in this field have been equally well-tested for their impact on democracy, rather than just their ability to collect and curate citizen inputs. Work such as (Cherry et al. 2021) showcases how citizens participating and being engaged in deliberative processes addressing specific issues tend to become more concerned and engaged about that issue through the process of engagement itself. The fact that AI platforms designed to facilitate citizen deliberation have been increasingly introduced in deliberative democracy practices (especially in its online versions) is therefore promising, but the ways in which AI has been introduced in them has been considered at times troubling for democratic ideals, seeking to replace rather than augment human decision-making. The many computer science experiments published in this space often reveal significant weaknesses in this respect, misunderstanding key characteristics of democratic exchange, casting attributes of discourse such as a lack of consensus, or questions of representation, as barriers to be removed, rather than essential aspects of the process of creating inclusive decisions. Scholars and practitioners have in fact documented how these technologies for online deliberation have been developed from a technocentric point of view and with little (or no) engagement or scrutiny by citizens, whose deliberative exercises these technologies are supposed to enhance. As Mathias Risse bluntly frames the challenge: 'Much as technology and democracy are not natural allies, technologists are not natural champions of or even obviously qualified advisers in democracy' (Risse 2023, p. 71).

Similarly, the policy-facing research of the past five years also seems to struggle to adapt effectively to the changing circumstances framing democratic technologies. For example, many research articles propose to discuss the tensions between democracy and AI without putting forward a clear definition of either of these two concepts. This lack of precision often groups multiple, varied phenomena being discussed as if they were the same, or proposes rough taxonomies of use without any structure or continuity between them. It also leads to unfortunate terminological simplifications, such as 'democracy by design' (Coeckelbergh 2024), to take hold, a discursive tendency which (Hadar et al. 2018) has compellingly demonstrated in the context of ‘privacy by design’ to represent a misplaced faith in the builders of technology to be able to take on stewardship of the social goods they propose to protect. A meta-analysis of this body of work finds the primary themes to include: ‘(1) the opportunities created by AI to improve democratic participation, (2) the challenges and ethical concerns that AI introduces, and (3) the ongoing efforts and necessary frameworks for democratic governance of AI technologies,’ (Sujana et al. 2025), a characterisation which points to the combination of technosolutionism and vagaries described above. The best work in this corpus, such as the article by Overton cited at the start of this section, manages to bring together a very clear contextual understanding of where and how democracy operates, and for whom, with precise and informed thinking about how technology intersects with these people in the context of this venue for democratic exchange. The inter- and transdisciplinary projects of the AI, Big Data and Democracy Forum have also focussed on identifying such clear angles on the wider landscape of complexity, with each working to build common ground between the values of participatory democracy (such as plurality, contextuality, complexity) and software design, which privileges instead clear categorisation, reduction of noise in data, and waterfall processing.

If we are to be able to protect the rights and values of European citizens within the important conduits of inclusive democratic exchange underpinned by big data and AI, we will need to come to a fundamental understanding of where the technologies and the needs of their users as citizens and members of communities and cultures, rather than primarily as technology consumers, come into conflict. Achieving such a goal will require discourse to be able to reach across boundaries of discipline, sector and practice without losing precision and focus on its objects and goals. Such an understanding will also unlock the potential for AI and big data to be used more widely for good in relation to these same rights, values and democratic processes, enabling expanded exploration of the potential of pro-social AI and big data to reinforce fundamental rights and European values, just as it underpins improved human health, more efficient services and greater access to information. Such a potential will be able to overturn the existing discourses, in which citizens’ weak interest and experience of the AI that might mediate their civic participation have been framed largely in negative terms: as a lack of interest, or an information deficit (Wilson 2022). Reversing this paradigm will also allow us to open up debates and policies around this topic, conversations which have often been happening behind closed doors and driven by technology business interests, resulting in biased and narrow ‘technological imaginaries’ (Ferrari 2020). A fundamental paradigm shift is required to overturn this discourse ‘which legitimates an established dominant industry, capable of influencing political actions’ (Ferrari 2020, p. 123). The examples that follow from the AI, Big Data and Democracy Forum projects point toward the directions in which this momentum might take us.

3 A Multi-stakeholder Empowerment. KT4D's Capacity Building and Regulatory Tools

The KT4D conceptual framework for approaching the challenges inherent at the interface of democracy and technology has been built, at its most fundamental level, upon two premises. First, that there is no way to approach this complex landscape from a single, unitary perspective. Indeed, we have determined, there are at least three sets of actors in constant balance to deliver a democracy-safe technological culture and society. This may seem obvious, but discursive trends demonstrate otherwise, with technology providers placing responsibility for patterns of technology use in the hands of users as responsible agents; users looking to policymakers to enstate regulatory approaches that will protect their democratic interests, and regulators seeking to balance the interests of the economy as a whole with the expanding corporate appetites for users and their data which are driving the technology sector. Just as no one of these voices can dominate, however, no one of them can be allowed to take a passive role, leading the KT4D project structure to itself around conjoined commitments to advance innovation, bringing pro-social technology developments to the fore, promote education, developing citizen capacities for critical digital literacy, and enhance regulation, as both a high-level value to underpin official policy, but also a more distributed critical capacity. We conceptualise this capacity around the idea of control, a term we seek to rehabilitate from a somewhat negatively-charged conceptual field that has been built up around it, as will be described a bit more below.

Generally, when cultural scholars think about control in the context of technology and society, their first point of reference is likely to be Gilles Deleuze's idea of the control society, as described in the 'Postscript on the Societies of Control' (1992). For Deleuze, the control society appeared as a further development of Foucault's 'environments of enclosure,' a 'new monster' (1992, p. 4) represented in societies dominated by computers, capitalism driven by a marketing imperative, and humans perpetually in debt. This nihilistic view of the technologised age is echoed across other Marxist responses to contemporary societal reconfigurations, with computing being characterised as central to far more than the digital communication and information practices related directly to them (Franklin 2015, p. xv). We have chosen, however, to resist this wholly negative view of technology and the incentive systems it creates so as to recognise the tri-partite relationship we describe as one rich in inherent potential for the different actors to influence each other, to enact a moderating force over self-interest, creating a system in which mutual controls can create balance. Control is expressed in an official sense in policymaking and regulation, true, but also in the control that the potential users and creators of technology might exert in their own interests, and in the interests of creating an overall techno-cultural system that is sustainable.

There are a number of manifestations and venues in which this concept of control can play out, into which the KT4D project needed to embed its innovation activities. These issues cross over between the positionalities of the individual and the collective, and between the disciplinary perspectives of psychology, philosophy, and cultural studies, among others. For that reason, the project placed significant emphasis on defining an evidence base that could adequately inform its activities within this reappropriated concept of control, and on delivering a well-informed, multiperspectival positioning for its innovation activities through its Social Risk Toolkit (available on the *KT4D website*, https://kt4democracy.eu). The eight overarching, thematic modules that inform this toolkit are as follows:

Module A: The question of free will and how individuals feel big data and AI may be exerting subtle forms of control over their opinions and desires;
Module B: The question of trust as a relational regulating mechanism toward actors relying on big data and AI and relationships mediated through them;
Module C: Cultural and historical perspectives on how individuals and collectives respond to changing knowledge technologies and the concomitant shifts in power and control structures this implies;
Module D: AI, big data, democracy and who controls user data and how;
Module E: AI, big data, democracy and control over individual freedom;
Module F: How enhanced critical digital literacy can place greater control in the hands of citizens in the context of their mediated civic and democratic participation;
Module G: The role of cultural specificity in civic participation, and the tensions this creates between user agency requirements and the culture-blindness of technology platforms;
Module H: Participatory Machine Learning (PML) and Algorithmic Accountability (AA) as software development paradigms for protecting civic and democratic participation within technology platforms.

These research modules frame an overall perspective on how a system in balance might be fostered, what it might currently lack, and where control functions may need to be enhanced. These perspectives were used to inform all of the activities of the project, in particular the development of the Digital Democracy Lab (see Sect. 5.5) and the assets supporting critical digital literacy (see Sect. 3.3). They also shaped the results of the two KT4D outputs most directly targeting the idea of control, namely the policy framework (fostering policy-based control from above) and the Social Computing Compass (fostering technology development featuring self-control and ethical guidance from the ground up).

The policy framework is based upon an understanding of the relationship between AI and big data systems and democracy as constructed according to six pillars, which vary in nature and importance across the lifecycle of technology implementation. The integration of advanced knowledge technologies like AI into the social fabric entails significant impacts for democratic systems. To understand these impacts, it is helpful to view democracy through six pillars: participation, freedom, equality, knowledge, transparency, and the rule of law (for different models of democracy see e.g., (Held 2006)). These pillars represent the values that sit at the heart of democracy, yet which AI might come to challenge for the better or worse.

AI technologies fundamentally reshape the pillar of participation by altering how citizens engage in political processes. Participation rests on the active involvement of citizens in governance and the responsiveness of government to their will. Yet, social media algorithms can be utilized to influence or manipulate public opinion by shaping the information citizens see (Shukla and Tripathi 2024). Furthermore, there is a systemic risk of concentrating power in the hands of those who design and develop these systems, which can limit the effectiveness of public input. Conversely, AI also offers opportunities to enhance civic participation by creating representative channels for deliberation and making complex policy information more understandable to citizens (Combaz et al. 2025). As for the freedom pillar, extensive monitoring, profiling, and surveillance enabled by AI can deter free expression, particularly in sensitive contexts like policing or border control (Ashraf 2020b). However, AI can also facilitate freedom of association by helping individuals connect with like-minded groups or by better tracking state power to increase accountability.

The pillar of equality, which requires the protection of fundamental rights, is also challenged by the disruptive nature of AI systems. AI risks inadvertently reinforcing existing societal biases and discrimination if trained on unrepresentative data, potentially leading to the underrepresentation of certain groups in decision-making (Kordzadeh and Ghasemaghaei 2022). There are also concerns regarding economic inequality and unemployment, as AI development is currently driven largely by the profit motives of private companies. On the other hand, AI can be used to promote substantive equality by identifying hidden biases or providing accommodations for disabled citizens. Regarding the knowledge pillar, democratic processes rely on a well-informed electorate and a healthy information ecosystem. AI influences this foundation by increasing reliance on chatbots, algorithmic echo chambers and boosting misinformation, thereby diminishing the epistemic agency of citizens (Coeckelbergh 2023). Yet, AI can also improve educational outcomes through personalized tutoring and help reconcile differing views by identifying consensus between parties.

The transparency pillar serves as the backbone of democratic accountability by enabling citizens to trust the democratic process. Yet, algorithmic decision-making is often criticized for being opaque and unexplainable. This ‘black-box’ problem makes it challenging for citizens to scrutinize or challenge decisions made by public administrations (de Fine Licht and de Fine Licht 2020). Furthermore, exploitative data collection practices often occur without individual awareness. Finally, under the rule of law, all individuals and institutions must be accountable to equally applied legislation. AI risks undermining this through electoral manipulation or by upsetting the balance of power if one branch of government centralizes decision-making through unchecked AI use. However, AI also holds the potential to make the court system more efficient and transparent while helping identify signs of government overreach (Köbis et al. 2022).

Focusing on the above pillars of democracy helps us see that the democratic impacts of AI are likely to manifest unevenly across the lifecycle of AI systems. This lifecycle consists of four primary stages: design, data, development, and deployment. While democratic impacts, such as discriminatory decisions or disinformation, often materialize later in the deployment stage, the decisions that enable these harms are frequently made much earlier. Mapping governance approaches across the AI lifecycle helps identify the most opportune points for intervention to safeguard democratic values.

The design phase is arguably the most critical for ensuring democratic alignment, as it is here that the system's objectives, rationale, and potential societal impacts are defined. This stage is especially vital in public sector procurement, where agencies may otherwise lack oversight of the actual development process. Key governance mechanisms at this stage include compute governance (Sastry et al. 2024), which regulates the access to computational resources needed for advanced AI, and participatory design requirements (Parthasarathy et al. 2024). Participation and freedom are the pillars most relevant to this stage, as design choices set the incentives for privacy-respecting systems and citizen engagement. In the data phase, the focus shifts to the management, sourcing, and quality of training data (Janssen et al. 2020). Unrepresentative or low-quality data is a primary cause of algorithmic discrimination, making this stage essential for upholding equality. Freedom is also heavily implicated here through privacy governance and copyright legislation, which protect individual anonymity and prevent the unauthorized exploitation of creative works (Bennett and Raab 2020).

The development phase involves the training, testing, and validation of the AI model to ensure it is robust and fair. This stage is particularly relevant for the pillar of equality, as it is where algorithmic biases are evaluated and mitigated (Gray et al. 2024). It is also the most critical stage for transparency through the implementation of explainability policies and external model audits. These audits provide impartial oversight to identify risks that internal development teams might overlook (Shevlane et al. 2023). Finally, the deployment stage is where the system is integrated into real-world environments. Because societal and democratic effects actualize here, it is often viewed as the most important phase for safeguarding democratic practices. Governance during deployment emphasizes monitoring, liability, and competition policies. The rule of law is most singularly focused on this stage, requiring clear legal accountability for harms and ensuring that individuals have effective remedies when their rights are violated. The knowledge pillar also relates strongly to deployment through oversight mechanisms and the use of competition legislation to prevent the monopolization of the AI ecosystem by large platforms (Narechania and Sitaraman 2024).

The realization that democratic impacts are distributed across the AI lifecycle suggests that mere regulation at the deployment stage is insufficient. A risk-based regulatory approach, like the EU's AI Act tends to assume that the use of AI is desirable and manageable through technical solutions, which can sideline fundamental questions about whether AI should be deployed in certain contexts at all (Kaminski 2023). Furthermore, such risk regulation can struggle to address the diffuse and abstract harms that characterize democratic risks. To truly safeguard democracy, governance must adopt an 'infrastructural lens' that addresses the concentration of power within the digital stack. If the digital infrastructure facilitating the modern public sphere is owned and operated by a few private platforms driven by profit incentives, democratic oversight becomes a challenge. This necessitates a shift toward digital public infrastructure (DPI)—openly accessible digital resources designed to serve the public interest (Mazzucato et al. 2024). This means focusing on creating shared AI infrastructure like open datasets, cloud compute and public algorithms. This infrastructural turn is intrinsically linked to digital sovereignty: the capacity of a society and public institutions to meaningfully govern critical AI infrastructures, data, and capabilities on their own terms.

Tackling the concentration of power in AI likely requires instruments beyond regulation, such as investments into a sovereign European AI stack. A prominent example is the proposal for a 'CERN for AI', representing a large-scale, pan-European research and computer infrastructure that pools expertise and resources (Petropoulos et al. 2025). By creating shared research and infrastructure capabilities, the EU can seek to ensure that the foundational layers of the AI ecosystem are aligned with democratic values from their inception. To operationalize this vision, the roadmap for democratic AI governance developed in KT4D proposes a set of policy actions clustered into domains. First, regulatory enforcement focuses on building capacity for the AI Office and national authorities to effectively enforce regulations like the AI Act, GDPR, DSA and DMA to prevent regulatory capture and safeguard democracy. Second, public AI infrastructure prioritizes the creation of shared European resources such as the Eurostack (Bria et al. 2025), to reduce structural dependencies on foreign technology. Third, investments and innovation aim to support competitive European AI companies and open-source solutions by leveraging public procurement, such as 'buy European' requirements. Fourth, AI literacy and participation focuses on education, involvement of civil society and citizens' assemblies to render AI policymaking more participatory. Finally, research and standards underpin these efforts by funding experimental, paradigm-shifting AI research and shaping international standards to reflect democratic values.

In the short term (2026–2028), the priority is enforcing existing regulation and building public AI infrastructure. This focuses on providing sufficient resources and training for the EU AI Office and national authorities to enforce the fundamental rights provisions of the AI Act. During this phase, public procurement and tax incentives should be used to encourage responsible AI development that meets high standards for transparency, privacy, and civil society participation. Simultaneously, the EU should invest into initiatives like the European Chips Act 2.0 and digital commons to reinforce digital sovereignty. In the mid-term (2029–2032), the goal is democratic adoption. As infrastructure matures, democratic safeguards must become embedded as a default into European AI systems. This phase might include new legislation on further strengthening AI literacy, participatory governance mechanisms and ensuring public oversight of regulatory sandboxes. Furthermore, trends across the AI value chain must be tracked continuously to mitigate oligopolistic tendencies and concentration of power.

Finally, in the long term (2033–2035), the roadmap envisions exercising AI sovereignty. At this point, public AI infrastructure and democratic practices converge, enabling Europe to govern its whole AI stack rather than remaining dependent on foreign technology giants. Policies should support anticipatory governance, incorporating systematic impact assessments and technology assessment practices within institutions. The goal is to move from mere AI literacy to using and governing these systems to strengthen democratic ways of life. As such, the roadmap seeks to guide policymakers in reinforcing democracy through the use of knowledge technologies. Reflecting on the mundane but fundamental ways systems like recommender engines shape democracy, it becomes clear that legislative tools alone are limited; a culture of democratic governance must be built into the entire AI lifecycle and its underlying infrastructure.

The Social Computing Compass takes a very different approach to the same overall challenge of bringing into balance the relationship between the democratic practices of individuals and the manner in which technologies designed to become a part of these processes are designed, built and deployed. Originally conceived as a canvas-style ethical assessment tool to enhance the cultural sensitivity of software designers and their outputs, in its final format, the Compass has been realised as an Interactive Digital Narrative (IDN) under the title ‘Red Team Mission: The Incident at Pine Valley High’ (http://nexusdigital-redmission.com). This shift in format was decided upon after an early participatory design session, in which the participants made clear their lack of knowledge of various canvas or checklist-style tools for ethical auditing, and indeed their lack of interest in such tools. Instead, they felt that they would be more likely to be able to confront the limits of their own knowledge and experiences (and therefore their own unconscious biases) through a non-competitive game format, where they could be informed through problem-solving and exploration. The resulting IDN places the user into the role of an IT consultant hired by a fictional company to assess a recent complaint regarding the company’s flagship product, a closed social network system designed to be deployed within schools. The user is invited to explore a range of materials documenting the design and deployment process, including platform records of the bullying case that caused the client to make their complaint, but also interviews with the company’s technical lead and the school’s data controller; user specification inputs, such as a survey of parents and students; perspectives from the student body and community and more. In all, a total of 13 different assets are available for the user to explore (though they are not required to look at all of them or any in particular, or review them in any order) before they make their decision as to which of the five departments they feel carries primary blame for the widespread dissatisfaction with the platform’s performance.

In the course of their exploration, the users are exposed to a perfect storm of contributing factors, with the ultimate fault lying in a complex cultural context and a technical implementation blind to its implications. For each of the five possible sources of the problem users are asked to choose from—Design, Technical Build, Leadership, Cultural Environment, and Data Control—the feedback they receive informs them that they are correct, but only partially. The feedback outlines what contributions were in fact made by the factor they chose, but also what the limits of the impact of that factor were. They are also given the chance to learn more about real world problems caused by their chosen cause, and to explore other possible responses.

In this way, the IDN realises its mission to recognise the shared nature of control, and the shared responsibility for ethical software design and deployment where it can impact community interactions, realising the KT4D vision for empowerment as a multi-stakeholder project, in which sustained dialogue will be required to ensure our technologies serve our democracies.

4 From Technologies to Guidelines. How AI4GOV Safeguards Human Rights and Democratic Values

AI-driven tools promise efficiency, scalability, and evidence-informed governance, however as mentioned also in previous Sections (see Chap. Why AI and Democracy?) their deployment in democratic contexts raises fundamental concerns regarding transparency, accountability, fairness, and the protection of fundamental rights. In policy-making environments, where decisions affect citizens collectively and often asymmetrically, the use of AI technologies cannot be treated as a purely technical matter. Instead, it requires an explicit and systematic alignment with democratic values, legal safeguards, and societal expectations. The AI4GOV project addresses this challenge by advancing a holistic approach that translates technological innovation into concrete governance guidelines, ensuring that AI systems used in public administration remain human-centric, trustworthy, and inclusive. In that context, the project is grounded in a socio-technical understanding of AI-driven governance, according to which democratic legitimacy emerges from the interaction of multiple societal actors, including public authorities, scientific and technical communities, technology providers, and civil society. Within this perspective, AI systems are not treated as neutral and completely technological artefacts, rather than as institutional mediators that shape relationships between knowledge production, decision-making authority, and public participation. The latter reveals the need that approaches and solutions towards safeguarding human rights and democratic values should not be designed and developed through the application of regulatory compliance or technical optimisation, but require design choices that embed accountability, inclusivity, and transparency across organisational, technical, and societal dimensions. Central to this approach is the recognition that democratic safeguards must be embedded throughout the entire AI lifecycle, from data collection and model design to deployment, evaluation, and revision. Rather than treating ethics and regulation as external constraints imposed after technical development, AI4Gov integrates them as constitutive elements of system design. This orientation is formalised specifically through the project's Holistic Regulatory Framework (HRF), as introduced in Sect. 7 and Chap. "Human-Centered AI for Democracy", which provides a comprehensive governance model for AI-based democratic systems. The framework is grounded in European legal and ethical standards, including the EU Charter of Fundamental Rights, the GDPR, and the EU AI Act, while also drawing on international guidelines for trustworthy AI. At this point, it should be noted that this framework restates high-level principles, as well as operationalises them through interrelated governance dimensions such as fairness, non-discrimination, human oversight, explainability, accountability, public engagement, and societal well-being. The development of this framework followed a mixed-methods and participatory process, combining literature review, expert consultation, and empirical fieldwork with underrepresented and vulnerable groups. The latter ensured that regulatory and ethical requirements reflect concrete experiences of bias, exclusion, and opacity in public services, rather than abstract normative assumptions. As a result, democratic values such as inclusivity, transparency, and participation are translated into actionable requirements that guide both organisational practices and technical implementation. The framework thus provides a normative anchor that enables AI4GOV to align technological innovation with social legitimacy and fundamental rights protection.

These normative commitments are further reinforced through AI4GOV's approach to data governance, which addresses one of the most critical sources of democratic risk in AI systems: the concentration and opaque use of data. AI-driven policy-making relies on large and heterogeneous datasets, often produced and managed by multiple institutions, raising concerns about data ownership, provenance, accountability, and power asymmetries between public authorities and citizens. The project responds to these challenges through a decentralised data governance model that prioritises citizen-centricity, traceability, and trust. This model represents a deliberate shift away from organisation-centric data control toward participatory and co-governed data ecosystems. Through the use of decentralised identifiers, verifiable credentials, and blockchain-based mechanisms, AI4GOV enables transparent tracking of data flows and decision processes while remaining compliant with privacy and data protection requirements. In that context, citizens are no longer positioned as passive data subjects but are empowered as active participants who can authenticate themselves, control the disclosure of their data, and verify how their information is used within AI-supported policy processes. The integration of self-sovereign identity mechanisms and citizen-facing tools enables individuals to engage with public services and policy-making infrastructures in a transparent and auditable manner, reinforcing principles of informational self-determination and procedural fairness. At the same time, AI4GOV introduces governance mechanisms that support collective oversight and shared responsibility in decision-making. Policy rules, evaluation criteria, and procedural changes can be proposed, debated, and validated through auditable processes that combine institutional authority with participatory input. From a democratic perspective, these design choices directly address concerns related to opacity, unaccountable automation, and exclusion, as they allow both institutions and citizens to scrutinise not only AI outputs but also the rules governing their generation. This approach does not treat decentralisation as an unqualified good or a self-evident design ideal. Decentralised architectures are instead subjected to systematic normative and institutional scrutiny, with attention to how they reconfigure power relations, responsibility, and vulnerability among public authorities, private intermediaries, and citizens. While decentralisation can reduce single points of failure and counter excessive data concentration, it may simultaneously introduce new forms of opacity, coordination failures, and governance fragmentation. These dynamics can complicate the allocation of legal responsibility, hinder effective supervision, and make it more difficult for individuals to identify who is accountable for harmful outcomes or rights violations. To mitigate these risks, regulatory safeguards and design constraints are embedded into decentralised systems from the outset, drawing on evolving European regulatory developments in data protection, digital services, and AI. Such safeguards include clear specification of roles and duties among actors, enforceable accountability mechanisms, and robust logging, audit, and traceability features that support both institutional oversight and external scrutiny. In addition, procedural guarantees, such as accessible avenues for contestation, human review, and effective redress, are handled as integral components of system design.

As highlighted so far, regulatory frameworks and data governance models establish essential safeguards, however their effectiveness ultimately depends on how they are embedded within technical architectures. A defining feature of the architecture is its alignment with the Quintuple Helix model, introduced in Sect. 7.4, which conceptualises innovation as a socio-technical continuum involving government, academia, industry, civil society, and the environment. The operationalisation of this model transforms AI-driven policy-making to a collaborative process going beyond technocratic exercises. Thus, AI4GOV's reference architecture plays a critical role in translating democratic principles into concrete technical capabilities, as depicted in the below Figure that represents the architecture blueprint of the project (Fig. 1).

The architecture is explicitly policy-aware and value-sensitive, ensuring that ethical, legal, and societal requirements are systematically mapped to software components, data flows, and system interactions. Explainability, bias mitigation, or human oversight are considered as foundational elements of AI-driven policy-making pipelines and are integrated into the architectural designs. This is reflected in the modular design of the platform, which includes dedicated components for bias detection and mitigation, explainable AI and situational awareness, policy recommendation, and interactive visualisation. Each component is linked to clearly defined user and technical requirements derived from stakeholder needs and pilot use cases, reinforcing traceability between societal concerns and technical solutions. Human-in-the-loop decision-making and interactive exploration of AI outputs are further supported to ensure that AI systems augment rather than replace human judgement, enabling policymakers, experts, and citizens to critically engage with algorithmic insights. The ultimate scope of the introduced architecture is not to function as an internal technical blueprint, but also as a translational instrument that bridges policy objectives and implementation practices. Explainability requirements articulated in the HRF are realised through integrated XAI libraries and self-explanatory visualisation tools, while accountability and auditability are supported through logging, monitoring, and provenance mechanisms embedded at infrastructural level. In this way, AI4GOV demonstrates how regulatory and ethical commitments can be operationalised through design, reducing the gap between normative intent and operational reality.

The overall alignment between regulatory frameworks, data governance mechanisms, and architectural design, showcases how projects, such as AI4GOV, ORBIS,

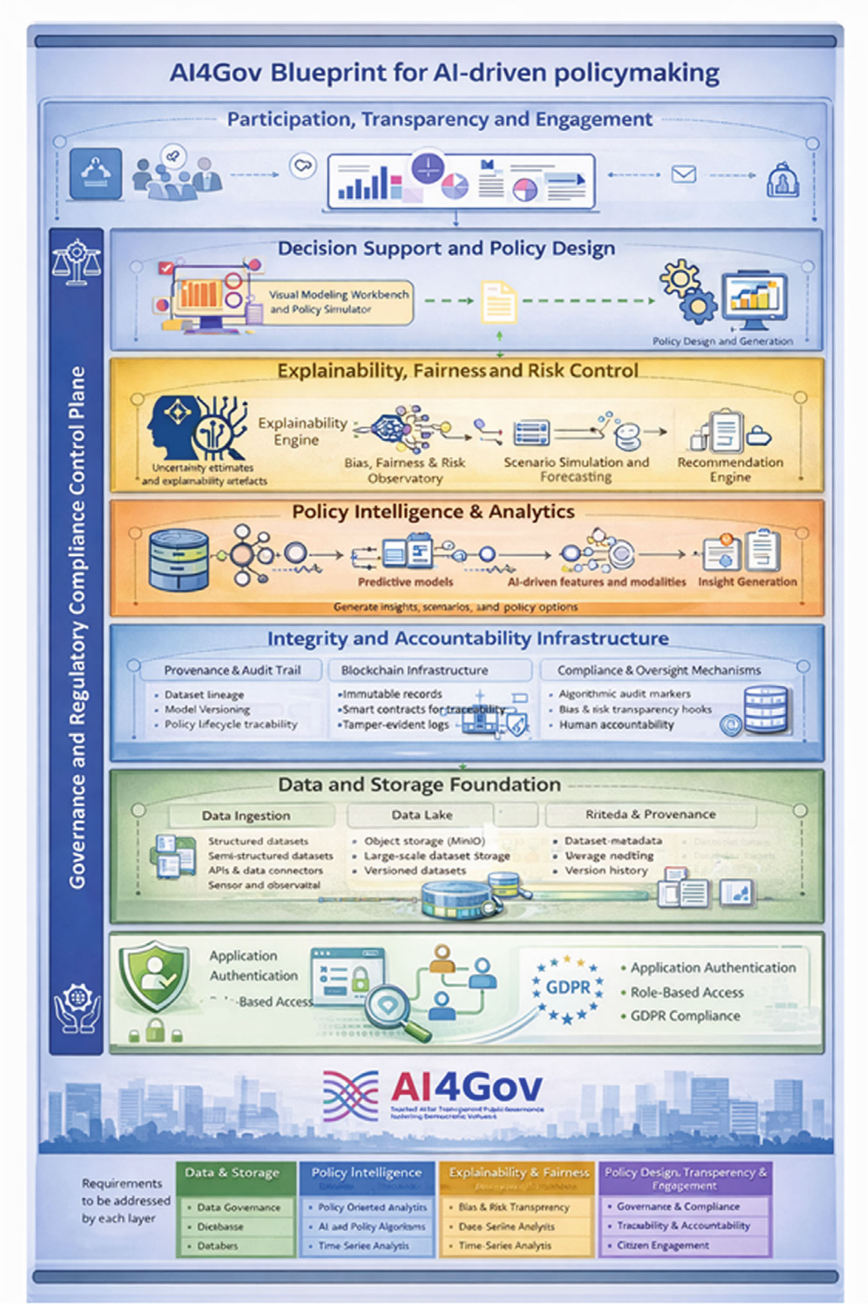

Fig. 1 The AI4Gov architecture blueprint

and ITHACA, as well as public administrations and institutions can move systematically from technological innovation to actionable governance guidelines. It illustrates that AI systems used in public administration should be policy-aware by design, with democratic values and legal safeguards explicitly translated into technical specifications. It highlights the importance of citizen-centric data governance as a prerequisite for trust, transparency, and legitimacy, particularly in contexts where AI informs collective decision-making. It also underscores that human oversight, explainability, and continuous monitoring are infrastructural necessities rather than optional enhancements in democratic governance. These key propositions contribute to a redefinition of innovation in democratic policy-making. They challenge the assumption that technological advancement and the protection of fundamental rights are inherently in tension, showing instead that democratic values can act as enabling conditions for sustainable and legitimate AI adoption. Hence, the embedding of socio-technical co-creation into design choices and translating it into concrete technological safeguards offers a transferable model for AI-enabled governance that prioritises trust, inclusivity, and societal benefit. The latter results in shaping a future vision of how AI can be applied into democratic processes in which technologies serve efficiently, accountable, collectively, and respecting the public good.

5 The Role of Evaluation to Look at Marginalisation. The ORBIS Framework

Marginalisation constitutes a central challenge for deliberative democracy and needs to be explicitly addressed when assessing the legitimacy and quality of deliberative processes. From a normative perspective, deliberative legitimacy rests on the capacity of all those affected by collective decisions to participate in processes of mutual justification under conditions of equality (Dryzek 2009; Karpowitz and Raphael 2016). Yet, a substantial body of critical scholarship has shown that marginalisation in deliberative settings cannot be understood merely as the absence of participation (Sanders 1997; Young 2002). Young's (2002) notion of internal exclusion captures situations in which participants are formally included but structurally disadvantaged in their ability to speak, be heard, or have their contributions taken seriously. Similarly, Sanders (1997) demonstrates that deliberation often presupposes communicative capacities—such as confidence, rhetorical skill, and familiarity with institutional or policy language—that are unevenly distributed across social groups. As a consequence, deliberative arenas may privilege already empowered actors while marginalising those whose communicative styles, experiences, or forms of expression diverge from dominant norms. More recently, these critiques have been reinforced through the concept of epistemic injustice (Fricker 2007). This perspective highlights how injustice arises when individuals or groups are systematically disadvantaged in their role as knowers, either through credibility deficits or through the exclusion of their interpretative frameworks. Applied to deliberative democracy, epistemic injustice draws attention to how marginalisation persists when certain forms of knowledge, particularly experiential or narrative, are undervalued or excluded from collective reasoning (Schmidt 2024).

In particular, marginalisation can occur at multiple stages and levels of deliberative processes. At the level of engagement, economic, linguistic, temporal, or digital barriers may shape who is able to participate in the first place (Silver et al. 2010; Pedrini et al. 2013). Even when access is formally ensured, interactional and discursive dynamics may generate internal exclusion through unequal speaking time, dominance by particular actors, or facilitation practices that fail to counterbalance power asymmetries (Sanders 1997; Young 2002). Marginalisation may also arise through issue framing and agenda-setting, which can privilege certain problem definitions while rendering others invisible. When deliberative agendas are framed in technocratic or abstract terms, the lived experiences of marginalised groups may struggle to gain recognition, resulting in epistemic marginalisation (Dryzek 2009; Schmidt 2024). Finally, marginalisation may persist at the level of outcomes and uptake, when deliberative outputs fail to influence policy or when participants, particularly those from disadvantaged groups, perceive their contributions as inconsequential (Hartz-Karp and Briand 2009; Gastil and Richards 2017). Such dynamics may foster disillusionment and mistrust towards deliberative processes and their claimed impacts.

Furthermore, the increasing use of digital platforms and AI-based tools in deliberative settings further complicates these challenges. While digital deliberation is often associated with promises of greater reach, scalability, and inclusiveness, empirical evidence suggests that these alternative settings frequently reproduce or reshape existing inequalities, linked to digital divides, differential engagement patterns, and platform-specific interaction dynamics (Kennedy et al. 2021). In this context, the risk of algorithmic epistemic marginalisation emerges, whereby minority positions or non-standardised forms of expression are distorted, deprioritised, or rendered less visible by automated systems (Schmidt 2024). These dynamics underscore the fundamental importance of continuous monitoring and of maintaining a sustained focus on how deliberative processes are designed and implemented to address marginalisation. Marginalisation cannot be treated as a static condition or a one-off problem to be solved at the point of access; rather, it emerges and evolves throughout the different stages of deliberation, shaped by design choices, facilitation practices, and technological mediation.

For this reason, evaluating deliberative processes is essential to making marginalisation visible and in enabling timely interventions. This need becomes even more pressing in deliberative processes supported by AI. AI-enhanced deliberation unfolds within socio-technical settings (Badham et al. 2000) shaped by the interaction of human actors, institutional arrangements, organisational structures, and technological infrastructures. AI systems that integrate computational and human components within shared deliberative environments are themselves socio-technical systems, and their deployment inevitably reshapes participation dynamics, influencing whose contributions are amplified or filtered, how authority and credibility are distributed, and which forms of knowledge gain visibility. As a result, marginalisation is not simply 'brought into' deliberation by participants' pre-existing disadvantages, but can be reinforced or mitigated by how deliberative environments are designed, governed, and technologically supported. Evaluating AI-enhanced deliberation cannot be confined to technical performance metrics, nor can it rely solely on abstract democratic ideals detached from technological mediation. Instead, evaluation frameworks should systematically examine the interplay between human actors, organisational structures, and AI systems, assessing how this interplay enables or constrains democratic practices, with particular attention to its effects on marginalised groups.

From this perspective, evaluation should be framed to move beyond a purely outcome-oriented exercise and become a systemic and reflexive practice. It should be designed and put in place to support the identification of marginalisation patterns across different stages of deliberation, capture interdependencies and feedback loops between social and technical components, and provide the empirical basis for adaptive design and governance choices over time.

It is within this conceptual and analytical context that the ORBIS project illustrates its contribution. Indeed, ORBIS conceptualises marginalisation not as a fixed attribute of individuals or groups, but as a socio-technical outcome of deliberative design choices and technological mediation. Building on this understanding, the ORBIS evaluation framework is designed to make marginalisation visible, comparable, and actionable in real-world deliberative processes, particularly those supported by AI-based tools. The following paragraphs introduce this framework, illustrating how its structure and analytical dimensions provide concrete pathways for identifying, assessing, and addressing marginalisation within deliberative processes.

5.1 The ORBIS Evaluation Framework

This paragraph introduces the ORBIS Evaluation Framework[1] as a concrete testbed for operationalising a socio-technical approach to evaluation in deliberative democracy, with a specific focus on identifying and addressing marginalisation. Rather than treating evaluation as a post hoc assessment, the framework is conceived as an integral component of deliberative design and governance. Its purpose is to systematically observe how both process design and technological mediation, particularly AI-based tools, shape the inclusiveness of deliberative contexts and influence whether participants are able to engage meaningfully and perceive themselves as active contributors to the debate.

[1] The ORBIS Evaluation Framework was developed by Nicos Stylianou (Center for Social Innovation), Ilaria Mariani (Politecnico di Milano, Department of Design), and Grazia Concilio (Politecnico di Milano, Department of Architecture and Urban Studies), as described in Deliverable D6.1 "*The ORBIS evaluation framework*", available at: https://orbis-project.eu/resources.

Within the ORBIS framework, deliberative democracy provides the normative foundation for evaluation. Deliberation is understood as a practice aimed at fostering mutual understanding, critical reflection, and transformative learning among diverse stakeholders. Accordingly, evaluation is not limited to measuring participation levels, but is explicitly oriented towards assessing inclusion, representation, discursive equity, and opportunities for meaningful engagement, with particular attention to groups at risk of marginalisation. At the same time, the framework draws on socio-technical studies to conceptualise deliberation itself as a socio-technical artefact. From this perspective, evaluation captures how AI tools and deliberative practices co-evolve over time, and how the introduction of technological mediation reshapes roles, power relations, interaction patterns, and infrastructural conditions within deliberative settings.

Responsible Research and Innovation (RRI) provides a further regulatory and ethical reference for the ORBIS evaluation framework (Burget et al. 2017; Crabu et al. 2022; Camarinha-Matos et al. 2023). RRI principles (anticipation, reflexivity, inclusion, and responsiveness) are operationalised as evaluative criteria that guide both the assessment of deliberative processes and the governance of technological innovation.

Taken together, these domains form the ORBIS socio-technical evaluation framework, which is both theoretically grounded and operationally applicable. ORBIS adopts a dual-lens evaluation design, combining a deliberative lens and a technological lens to examine how deliberative quality and technological mediation interact and co-evolve across the different stages of deliberation, and how this interaction may generate, reproduce, or mitigate forms of marginalisation affecting participation and meaningful engagement in deliberative democracy.

The Deliberation Lens

The Deliberation Lens focuses on the democratic quality of deliberative processes and examines whether participatory practices are inclusive, adaptable, and meaningful within public-sector decision-making. The lens is structured around four interrelated dimensions.

(1) **Methods** examine whether deliberative approaches are flexible, scalable, and adaptable to different institutional, cultural, and policy contexts. Evaluation goes beyond asking whether a method ‘works’ in a single setting, and instead considers whether it can be transferred or adjusted without undermining legitimacy, inclusiveness, or deliberative quality.
(2) **Inclusion and representation** focus on who takes part in deliberation and how. This dimension assesses whether marginalised or disadvantaged groups are meaningfully involved and whether participants have genuinely equal opportunities to contribute, to be heard, and to influence outcomes.
(3) **Topics** address what is deliberated. Here, evaluation considers the diversity and relevance of issues under discussion, as well as their alignment with participants’ interests, lived experiences, and expectations, and with the stated aims of the deliberative process.

(4) **Results** capture both immediate and longer-term effects of deliberation. This includes the feasibility and sustainability of recommendations, participants' willingness to remain engaged over time, and the extent to which deliberation contributes to building trust, strengthening relationships, and fostering ongoing collaboration among stakeholders.

The Technology Lens

Complementing the deliberative perspective, the Technology Lens examines how digital and AI-based tools shape deliberative practices. This lens asks whether technological mediation actively supports democratic interaction, reason-giving, and mutual understanding, or whether it introduces new barriers and asymmetries. This lens is articulated around four main dimensions.

(1) **Interaction** considers usability, accessibility, and the quality of feedback mechanisms, assessing whether digital interfaces genuinely support participation for users with different abilities, levels of confidence, and technological literacy.
(2) **Argument management** examines whether systems help participants articulate, relate, and contest arguments in constructive ways. The focus is on whether AI tools support meaningful disagreement and understanding, or whether they risk oversimplifying, misrepresenting, or distorting participants' contributions.
(3) **Cross-platform accessibility** addresses compatibility across operating systems and levels of digital competence. This dimension explicitly considers digital divides and evaluates whether technological choices exclude certain groups or limit their ability to participate fully.
(4) **Visualisation** looks at how information, arguments, and outcomes are shared and made accessible. Evaluation focuses on whether visual representations are accessible, interpretable, meaningful, and useful for diverse participants.

Measurement Tools

To translate this dual-lens framework into practice, ORBIS employs eight complementary measurement tools, each capturing different aspects of AI-enhanced deliberative processes.

(1) **Implementation plans** provide structured roadmaps for pilot activities and enable systematic monitoring over time.
(2) **Self-assessment questionnaires** invite pilot organisations to reflect on methods, tools, and design choices.
(3) **Pilot diaries** offer qualitative, longitudinal insights into how deliberative processes evolve in practice, documenting challenges, adaptations, and learning.
(4) **Platform dashboards** generate data on user experience and interaction patterns.
(5) **User registration and profiling** help identify who is reached.
(6) **User questionnaires** capture participants' perceptions of inclusion, engagement, and satisfaction.
(7) **Expert interviews** provide contextualised, interpretative insights.

(8) **Foresight scenarios** explore longer-term implications, scalability, and potential systemic effects.

Used together, these instruments enable triangulation, allowing evaluators to compare perspectives, validate findings, and build a richer understanding of how marginalisation emerges across socio-technical contexts.

Levels of Analysis

The ORBIS evaluation framework operates across three interconnected levels of analysis, reflecting the multi-scalar nature of AI-enhanced deliberation.

- At the **participant level**, evaluation focuses on individual experiences of inclusion, accessibility, learning, and motivation. Key questions include whether participants feel listened to, whether deliberation supports learning and reflection, and whether individuals are willing to remain engaged over time.
- At the **process level**, evaluation examines deliberative methods and technological infrastructures, assessing their adaptability, their capacity to support inclusive interaction, and the alignment between topics, tools, and participants' expectations.
- At the **ecosystem level**, the framework considers broader systemic effects, including policy influence, the emergence of new partnerships or networks, and contributions to institutional learning and democratic innovation.

This multi-level approach enables a holistic evaluation that connects micro-level experiences with meso-level process dynamics and macro-level institutional outcomes, reflecting the assumption that AI-enhanced deliberation operates simultaneously across these scales and that evaluation must be capable of capturing this complexity in a context-sensitive and non-prescriptive way. On this ground, the ORBIS Evaluation Framework organises its aforementioned components (the two lenses, the levels of analysis, and the associated measurement tools) within a structured evaluation matrix aimed at providing a systematic representation of the framework, detailing how normative principles and evaluative goals are translated into operational indicators. Each row of the table corresponds to a specific performance indicator and is organised according to a consistent set of fields, which collectively describe both the conceptual rationale and the practical implementation of evaluation.

Specifically, the evaluation matrix is structured around the following elements: the Lens to which the indicator belongs; a numerical identifier (N.) and Dimension; the Sub-dimension and its descriptive rationale; the Performance Indicator and its descriptor; the Target Group addressed; the Targeted KPIs; the theoretical References grounding the indicator; the Measurement Tools used for data collection; the associated Guiding Questions; and the overarching Research Questions to which the indicator contributes.

5.2 *Marginalisation in the ORBIS Evaluation Framework: The Deliberation Dimension*

ORBIS embeds the analysis of marginalisation primarily within the **Deliberation Lens** (Table 1).

At the level of **methods**, indicator **D01.01** focuses on participant diversity across demographics and backgrounds, requiring each pilot to define marginalised groups contextually and to assess whether these groups are substantively represented. This approach is grounded in work showing that diversity in participation is a prerequisite for legitimacy and long-term policy relevance, particularly in complex or contested policy domains (Abdel-Monem et al. 2010; Wang et al. 2015). Diversity here is therefore treated as a proxy for the capacity of deliberative methods to surface perspectives that would otherwise remain peripheral. Complementing this, **D01.06** examines the alignment of deliberative methods with participants' expectations and needs. This indicator reflects findings from participatory governance research emphasising that perceived procedural fit and responsiveness are critical for sustaining engagement, especially among participants with lower levels of policy literacy or deliberative confidence (Abdel-Monem et al. 2010).

The second cluster of indicators explicitly targets **inclusion and representation**, addressing marginalisation more directly. **D02.01** evaluates whether marginalised groups—defined in relation to each pilot's objectives and context—are intentionally identified and meaningfully included. This indicator reflects normative arguments in deliberative theory that inclusion should be context-sensitive and attentive to structural inequalities rather than based on fixed or universal categories (Miller 2003; Deveaux 2018). **D02.02** complements this qualitative assessment by examining the proportion of participants from disadvantaged backgrounds. Empirical studies on participatory processes highlight that numerical presence remains an important diagnostic signal for identifying structural barriers to access, even when qualitative inclusion strategies are in place (Quick and Feldman 2011; Pedrini et al. 2013). Within ORBIS, this indicator supports reflection on recruitment strategies, access conditions, and support mechanisms rather than functioning as a rigid benchmark. **D02.03** focuses on equal opportunities for contribution, capturing whether all participants perceive that they can meaningfully intervene in deliberation. This indicator is informed by research showing that discursive inequalities can persist within formally inclusive settings, affecting both perceived fairness and deliberative quality (Chaskin and Greenberg 2015; Velikanov 2017). Similarly, **D02.04** addresses socioeconomic diversity, recognising that income level, occupational background, and related constraints shape not only who participates, but also how participants engage and how legitimate or risky they perceive participation to be. This dimension draws on evidence that socioeconomic position influences deliberative behaviour, confidence, and patterns of voice in participatory settings (Silver et al. 2010; Vargas et al. 2017).

Marginalisation can also occur through **topic selection and framing**. Indicator **D03.01** therefore assesses the variety of perspectives represented across discussions.

Table 1 The deliberation lens matrix

Dimension	ID	Performance indicator	Descriptor of indicator	Guiding questions
Methods	D01.01	Participant diversity across demographics and backgrounds	Measures the extent to which deliberative methods succeed in engaging a broad and diverse range of participants, including marginalised groups defined contextually	How well do participant demographics reflect the diversity required to meet the pilot's objectives?
Methods	D01.06	Participants' expectations were met	Assesses whether deliberative methods align with participants' needs, expectations, and levels of expertise	To what extent did the methods align with participants' expectations and needs?
Inclusion and representation	D02.01	Participant diversity by inclusion categories	Evaluates whether marginalised groups are intentionally identified and meaningfully included in the deliberative process	Which marginalised groups were included, and how effectively were they supported during deliberation?
Inclusion and representation	D02.02	Percentage of participants from disadvantaged backgrounds	Measures the representation of participants from disadvantaged or underrepresented backgrounds	How many participants from disadvantaged backgrounds were included, and what barriers affected their participation?
Inclusion and representation	D02.03	Equal opportunities for contribution	Assesses whether all participants perceive that they have fair and equal opportunities to contribute to deliberation	How equitable were opportunities for different participants to contribute during deliberation?
Inclusion and representation	D02.04	Socioeconomic diversity	Captures the inclusion of participants from diverse income levels and occupational backgrounds	Were participants from varying socioeconomic backgrounds able to participate fully and comfortably?
Topics	D03.01	Variety of perspectives represented	Measures the presence of diverse and contrasting viewpoints within deliberative discussions	To what extent did the discussion topics enable the inclusion of diverse perspectives?

(continued)

Table 1 (continued)

Dimension	ID	Performance indicator	Descriptor of indicator	Guiding questions
Outcomes	D04.06	Influence of participants on deliberation results	Assesses whether participants are empowered to meaningfully influence deliberative outputs	To what extent were participants able to influence deliberative outcomes?
Outcomes	D04.07	Citizen-driven innovation actions and partnerships	Evaluates whether deliberation stimulates citizen-led initiatives, collaborations, or follow-up actions	Did deliberation lead to citizen-driven innovation actions or new partnerships?

This indicator builds on deliberative scholarship emphasising that meaningful deliberation depends on the visibility and recognition of plural and potentially conflicting viewpoints (Dryzek 2009, p. 200; Karpowitz and Raphael 2016).

Finally, ORBIS considers marginalisation in relation to **outcomes**. Indicators **D04.06** and **D04.07** assess whether participants are empowered to influence deliberative outputs and whether deliberation stimulates citizen-driven innovation, partnerships, or follow-up actions (Papadopoulos and Warin 2007; Hartz-Karp and Briand 2009; Delina 2018).

These indicators operationalise marginalisation as a **multi-dimensional and socio-technical phenomenon**, emerging from the interaction between methodological design, social dynamics, topic framing, and institutional impact.

5.3 Marginalisation in the ORBIS Evaluation Framework: The Technology Dimension

In relation to marginalisation, the **Technology Lens** plays a crucial role in unpacking data and identifying whether digital and AI-based tools effectively meet the needs of diverse and potentially marginalised participants (Table 2). The Technology Lens complements the Deliberation Lens by providing empirical evidence on how inclusion is practically enacted or constrained through technological design and use.

Indicators related to **interaction** (**T01.01**) assess the capacity of the ORBIS toolkit to engage a broad and diverse audience digitally, using registration data as a proxy for outreach and accessibility. Prior studies have shown that levels of digital engagement can signal both inclusionary potential and structural barriers related to access and motivation (Bojic et al. 2016; Shin et al. 2022). High registration numbers alone do not guarantee inclusion; however, when read alongside participant profiling and deliberation indicators, they allow evaluators to identify which groups are reached through digital channels and which remain underrepresented. This mapping function is particularly relevant for detecting structural exclusion related to digital divides.

Table 2 The technology lens matrix

Dimension	ID	Performance indicator	Descriptor of indicator	Guiding questions
Interaction	T01.01	Number of registrations	Indicates the capacity of the digital toolkit to engage a broad and diverse audience and provides a proxy for digital access and outreach	Who is reached through digital registration channels? Which groups remain underrepresented?
Interaction	T01.02	Usability of the ORBIS toolkit	Assesses how intuitive and user-friendly the toolkit is for participants and organisers with different levels of technological familiarity	How intuitive and easy to use is the toolkit for managing discussions and arguments?
Interaction	T01.03	Feedback and engagement mechanisms	Evaluates whether feedback features support sustained and meaningful engagement throughout deliberation	To what extent do feedback features encourage active and meaningful participation?
Interaction	T01.04	Technological understanding	Measures organisers' awareness and familiarity with the technology, reflecting trust and capacity for inclusive deployment	What is the level of familiarity with the ORBIS technology among organisers?
Interaction	T01.05	Iterative impact of toolkit outputs	Assesses whether toolkit outputs inform future planning and iterative improvement of deliberative processes	To what extent do toolkit outputs guide future deliberative planning and actions?
Argument management	T02.01	Conflict and disagreement management	Evaluates whether the toolkit supports constructive disagreement and conflict resolution	To what extent does the ORBIS toolkit support the prevention or resolution of conflicts?
Argument management	T02.02	Quality and representativeness of argumentation outputs	Assesses whether AI-generated summaries and arguments accurately reflect participants' contributions	Do the toolkit outputs accurately and fairly represent the diversity of participant inputs?

(continued)

Table 2 (continued)

Dimension	ID	Performance indicator	Descriptor of indicator	Guiding questions
Multiplatform accessibility	T03.01	Impact of the toolkit on deliberation	Measures the perceived contribution of the toolkit to the quality of deliberative processes	How effective is the toolkit in supporting deliberative debates and outcomes?
Multiplatform accessibility	T03.02	Integration with existing systems	Evaluates the alignment of the toolkit with systems already in use by organisers	To what extent is the ORBIS toolkit compatible with existing systems and practices?
Multiplatform accessibility	T03.03	Accessibility across devices	Assesses usability across desktop, mobile, and tablet devices to ensure inclusive access	Is the toolkit usable across different devices and technological contexts?
Visualisation	T04.01	Clarity of data visualisation	Evaluates whether visual outputs are clear, meaningful, and interpretable for diverse users	To what extent are deliberative outputs clearly and meaningfully visualised?

Furthermore, marginalisation in AI-enhanced deliberation often manifests through differential ability to interact meaningfully with digital tools. Research on decision-support systems and collaborative platforms highlights usability and intuitiveness as key preconditions for meaningful engagement and trust (Hansson et al. 2013; Iandoli et al. 2018). Effective feedback mechanisms have been shown to sustain engagement and foster deliberative continuity, particularly in digitally mediated discussions (Gastil and Richards 2017; Shin et al. 2022). Prior work emphasises that trust in socio-technical systems is closely linked to users' understanding of underlying technologies (Hansson et al. 2013; Bojic et al. 2016). Indicators addressing ease of interaction, feedback mechanisms, and technological understanding (**T01.02**, **T01.03**, **T01.04**) explicitly capture this dimension.

Beyond access and interaction, the Technology Lens addresses epistemic dimensions of marginalisation through indicators related to **argument management** (**T02.01**, **T02.02**). These indicators assess whether AI-supported functions accurately and fairly represent participants' contributions. Studies on digitally mediated deliberation underline the importance of technological support for constructive disagreement and conflict resolution, while cautioning against distortive effects of poorly designed mediation tools (Gastil and Richards 2017; Perez 2020). From a marginalisation perspective, this evaluation is critical. If algorithmic summarisation privileges dominant frames, simplifies minority positions, or misrepresents nuanced arguments, epistemic exclusion may occur even in otherwise inclusive settings. ORBIS therefore evaluates the perceived accuracy, representativeness, and usefulness of argumentation outputs as a way to assess whether AI tools support discursive justice and high-quality reasoning, rather than reinforcing existing asymmetries (Perez 2020).

Indicators within the **multiplatform dimension** (**T03.01**, **T03.02**, **T03.03**) further contribute to mapping marginalisation by examining the adaptability of the ORBIS toolkit to diverse technical environments. Research on participatory technologies highlights that compatibility with existing systems and accessibility across devices and existing tools are crucial for ensuring inclusivity in heterogeneous organisational and infrastructural contexts (Maciel and Bicharra Garcia 2007; Kučera et al. 2020).

Finally, indicators related to **visualisation** (**T04.01**) address the interpretability of deliberative outputs. Prior research shows that clear and well-designed visualisations are essential for supporting understanding, transparency, and informed engagement with complex data (Hullman and Diakopoulos 2011; Kennedy et al. 2016). From a socio-technical perspective, inadequate or opaque visualisation may disproportionately disadvantage participants with lower levels of technical or data literacy, reinforcing asymmetries in understanding and influence. ORBIS therefore evaluates visualisation as a democratic affordance supporting accountability and the meaningful appropriation of deliberative results.

The indicators of the Technology Lens allow ORBIS to **map marginalisation across the socio-technical configuration of deliberation**. By unpacking data on access, interaction, argument management, adaptability, and interpretability, the framework assesses whether technology functions as an inclusive infrastructure or as a site where new forms of exclusion may emerge. Technology-related indicators are systematically read in conjunction with those of the Deliberation Lens, enabling evaluators to trace how technological affordances interact with participation patterns, discursive dynamics, and outcomes. In doing so, ORBIS positions evaluation as a mechanism for identifying socio-technical mismatches and for informing adaptive design choices aimed at mitigating marginalisation.

5.4 Responsible Research and Innovation in ORBIS

The integration of **Responsible Research and Innovation** (**RRI**) within the ORBIS Evaluation Framework provides an explicit normative and regulatory reference for addressing marginalisation in AI-enhanced deliberative processes. RRI offers a framework to ensure that research and innovation are socially desirable and aligned with societal needs and values (von Schomberg 2013; European Commission 2014; Egeland et al. 2019).

Originally conceptualised by von Schomberg (2011) as a transparent and interactive process, RRI emphasises the shared responsibility of innovators, users, policymakers, and wider society in shaping socio-technical innovation.

Subsequent contributions have consolidated RRI around four core principles (anticipation, reflexivity, inclusion, and responsiveness (Stilgoe et al. 2013)) which together provide a normative compass for evaluating the societal implications of innovation. Anticipation calls for systematic reflection on potential impacts and unintended consequences; reflexivity requires actors to critically examine their own assumptions, values, and power positions; inclusion stresses the involvement of diverse perspectives, particularly those traditionally excluded; and responsiveness emphasises the capacity to adapt innovation trajectories in response to emerging concerns and feedback. These principles are particularly relevant for AI-enhanced deliberation, where technological mediation can both enable participation and introduce new forms of exclusion or epistemic asymmetry.

Recognising its growing importance, RRI has been progressively embedded within European research and innovation policy frameworks (European Commission et al. 2020). Nevertheless, defining what counts as 'socially desirable' innovation remains inherently challenging in pluralistic societies characterised by diverse values, interests, and expectations (Weckert et al. 2016). This tension becomes particularly salient in deliberative democratic contexts, where inclusion, representation, and fairness are not abstract normative ideals but concrete design and governance challenges.

Despite its substantial theoretical richness, RRI has therefore faced persistent difficulties in operationalisation. A recurrent concern in the literature is the gap between high-level normative principles and their translation into concrete practices, especially with regard to meaningful stakeholder engagement and mechanisms of evaluative accountability (Owen et al. 2012; Emery et al. 2014; Rip 2014). When not supported by clearly defined methods, indicators, and tools, RRI risks being perceived as vague, aspirational, or disconnected from everyday research and innovation practices (de Saille 2015; Burget et al. 2017).

The ORBIS project explicitly addresses these challenges by embedding RRI principles within its evaluation framework.

(1) **Anticipation** is reflected in the framework's capacity to identify potential exclusionary effects of both deliberative design and technological mediation. Through monitoring and the orientation of digital tools, ORBIS evaluation supports early detection of marginalisation risks, including those related to access, usability, and representation.
(2) **Reflexivity** is embedded through self-assessment instruments, pilot diaries, and expert interviews, which encourage organisers to critically reflect on methodological choices, power dynamics, and assumptions about participants' capacities and needs. This reflexive dimension is essential for uncovering subtle or unintended forms of marginalisation that may not be immediately visible through quantitative indicators alone.
(3) **Inclusion** is operationalised as a measurable dimension of evaluation. ORBIS explicitly assesses who participates, whose perspectives are represented, how contributions are mediated, and whether participants perceive equal opportunities to influence outcomes. In doing so, the framework aligns RRI's inclusive ambition with concrete participation-related indicators.

(4) **Responsiveness** is ensured through the iterative and adaptive nature of the evaluation process. Feedback from participants and organisers is systematically integrated into the refinement of deliberative methods and technological configurations, allowing ORBIS to respond dynamically to emerging needs, concerns, and contextual constraints.

The operationalisation of RRI principles within ORBIS is achieved through a dual-lens evaluation architecture that explicitly targets the detection and mitigation of marginalisation in AI-enhanced deliberative processes. The Deliberation Lens addresses marginalisation by assessing who participates, how marginalised groups are represented, whether discursive dynamics are equitable, and to what extent participants, particularly those at risk of exclusion, can influence deliberative outcomes. In parallel, the Technology Lens focuses on anticipatory and epistemic forms of marginalisation by evaluating how digital and AI-based tools enable or constrain access, interaction, argumentation, and the interpretability of deliberative outputs.

Beyond this dual perspective, ORBIS applies RRI principles across multiple levels of analysis, recognising that marginalisation manifests differently across scales. At the participant level, evaluation captures lived experiences of inclusion, accessibility, learning, and agency, making visible subtle and internal forms of exclusion. At the process level, it examines how deliberative designs and technological infrastructures may reproduce or mitigate unequal participation and discursive power. At the ecosystem level, it assesses whether deliberative processes contribute to or challenge structural forms of marginalisation through policy uptake, institutional learning, and the creation of inclusive governance arrangements.

By combining a dual-lens architecture with a multilevel analytical approach, ORBIS transforms RRI from a normative aspiration into a measurable evaluative practice that treats marginalisation as a socio-technical phenomenon. Evaluation thus functions both as a detection mechanism, identifying exclusionary dynamics as they emerge, and as an activation mechanism, enabling reflexive learning and adaptive redesign aimed at mitigating marginalisation throughout the deliberative process.

6 Conclusions

As the above chapter has demonstrated, the challenge of ensuring technologies deployed into democratic contexts maintain the value of inclusivity is a large one, for a number of reasons. The first of these has to do with technological design and data. AI and big data systems are notoriously closed and opaque, and corporate approaches to the collection of training and testing data have not always respected human values and democratic principles. The ORBIS project's approach to responsible research and innovation, embedding measures to resist exclusion into their systems at both the level of technological function and human auditing, demonstrates the scale of commitment necessary to offset the gravity of system and data bias so pervasive in software development. But it takes more than just a commitment to inclusivity to make sure that users are safeguarded from the tendencies of AI systems to flatten, simplify, and remove troublesome outliers falling beyond the standard parameters of a statistical model. Preserving inclusivity requires as well that complexity be embraced and transformed from a distraction to efficiency into an asset for democracy. For example, in pursuing this goal, both KT4D and AI4Gov discovered that lifecycle models were necessary additions to their conceptualisations of AI system regulation. Without a firm embedding of how the risks inherent in a system change over the duration of its design and deployment, it is nearly impossible not to conflate these constraints into larger entities that ultimately represent nothing well, and can therefore easily be disregarded. Similarly, all of the projects discussed in this chapter propose mechanisms by which user positionality and the numerous possible definitions of a system's successful function might be incorporated into an inclusive approach to system assessment. As ORBIS reminds us, quality of deliberation and of technology need not be in conflict, but they may be in tension, or not be expressible according to the same terminologies. This need not be a zero-sum calculation, however, but rather a representation of how controls (as KT4D would understand them) and safeguards (as AI4Gov would frame them) might exist in a system to ensure perspectives do not become marginalised, but rather checked against, balanced among, and integrated with each other.

Perhaps as a result of this commitment to productive decoupling of the contributors to, processing within, and outputs of an AI system, including its socio-technical context, all of the projects featured in this chapter also highlight a need to carefully observe ecosystems and processes. One cannot simply ignore some contributors within the Quintuple Helix model at certain times simply because this makes the object easier to bring into focus. These actors will always be found in a state of dynamic interplay, and it has been one of the great achievements of KT4D, AI4Gov and ORBIS that they have managed to find ways of taking snapshots of this moving landscape, and of acting within this ever-changing landscape of processes in motion, rather than discrete products. Only by addressing the complexity of democratic process, with and without the additional layer of AI processing, can we seek to promote and protect the value of inclusivity, safeguard human rights, and mitigate the gravity of forces leading to marginalisation.

The reason why inclusivity can be such a challenge has to do with policies and principles, and the way in which we may or may not create a common understanding not just of how these systems should be designed, but the constraints we need to place on how they are used, with and for whom. This is a context in which KT4D's conceptual centering of a revitalised, multiperspectival concept of control provides a powerful perspective from which to observe and assess technology adoption cycles. It integrates the approaches to promoting human agency that appear across the projects of the consortium, and the shared commitment to ensuring this value can become actionable in software systems. This is in particular the case in AI4Gov, where values pertaining to human rights have been realised within the context of their bias detection work to frame technology development so as to ensure protection and voice for all of its users. Bringing together regulation, education and innovation as mutually supportive, but also mutually balancing, forces allows responsibilities for democracy and the technologies that support it be shared appropriately among the parties that design, build, regulate and use it.

Funding 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

Abdel-Monem T, Bingham S, Marincic J, Tomkins A (2010) Deliberation and diversity: Perceptions of small group discussions by race and ethnicity. Small Group Res 41:746–776. https://doi.org/10.1177/1046496410377359

Ashraf BN (2020a) Devastation caused by COVID-19: is democracy to blame? SSRN 3596009. https://doi.org/10.2139/ssrn.3596009

Ashraf C (2020b) Artificial intelligence and the rights to assembly and association. J Cyber Policy 5:163–179. https://doi.org/10.1080/23738871.2020.1778760

Badham R, Clegg C, Wall T (2000) Socio-technical theory. In: Handbook of ergonomics. John Wiley New York, NY

Bennett CJ, Raab CD (2020) Revisiting the governance of privacy: Contemporary policy instruments in global perspective. Regul Gov 14:447–464. https://doi.org/10.1111/rego.12222

Bojic I, Marra G, Naydenova V (2016) Online tools for public engagement: case studies from Reykjavik. Preprint, arXiv, November 28. https://doi.org/10.48550/arXiv.1611.08981

Bria F, Timmers P, Gernone F (2025) EuroStack—a European alternative for digital sovereignty. https://www.bertelsmann-stiftung.de/en/publications/publication/did/eurostack-a-european-alternative-for-digital-sovereignty

Burget M, Bardone E, Pedaste M (2017) Definitions and conceptual dimensions of responsible research and innovation: a literature review. Sci Eng Ethics 23:1–19. https://doi.org/10.1007/s11948-016-9782-1

Camarinha-Matos LM, Ferrada F, Oliveira AI (2023) Implementing RRI in a research and innovation ecosystem. In: González-Esteban E, Feenstra RA, Camarinha-Matos LM (eds) Ethics and responsible research and innovation in practice: the ETHNA system project. Springer Nature Switzerland, Cham, pp 87–108

Chaskin RJ, Greenberg DM (2015) Between public and private action: neighborhood organizations and local governance. Nonprofit Volunt Sect Q 44:248–267. https://doi.org/10.1177/0899764013510407

Cherry C, Capstick S, Demski C et al (2021) Citizens' climate assemblies: understanding public deliberation for climate policy. Cardiff, July 31. https://orca.cardiff.ac.uk/id/eprint/145771/

Coeckelbergh M (2023) Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. AI Ethics 3:1341–1350. https://doi.org/10.1007/s43681-022-00239-4

Coeckelbergh M (2024) Why AI undermines democracy and what to do about it. John Wiley & Sons

Combaz A, Mas D, Sanders N, Victor M (2025) Applications of artificial intelligence tools to enhance legislative engagement. https://www.alphaxiv.org/abs/2503.04769

Crabu S, Mariani I, Schmittinger F (2022) Framing real-life experimentations as case studies. In: Deserti A, Real M, Schmittinger F (eds) Co-creation for responsible research and innovation: experimenting with design methods and tools. Springer International, Cham, Switzerland, pp 25–35. https://doi.org/10.1007/978-3-030-78733-2_3

de Saille S (2015) Innovating innovation policy: the emergence of 'Responsible Research and Innovation.' J Responsible Innov 2:152–168. https://doi.org/10.1080/23299460.2015.1045280

de Fine LK, de Fine LJ (2020) Artificial intelligence, transparency, and public decision-making. AI Soc 35:917–926. https://doi.org/10.1007/s00146-020-00960-w

Deleuze G (1992, October) Postscript on the societies of control, vol 59, pp 3–7

Delina LL (2018) Energy democracy in a continuum: remaking public engagement on energy transitions in Thailand. Energy Res Soc Sci 42:53–60. https://doi.org/10.1016/j.erss.2018.03.008

Deveaux M (2018) Deliberative democracy and multiculturalism. In: The Oxford handbook of deliberative democracy, pp 156–170

Dryzek JS (2009) Democratization as deliberative capacity building. Comp Pol Stud 42:1379–1402. https://doi.org/10.1177/0010414009332129

Egeland C, Forsberg E-M, Maximova-Mentzoni T (2019) RRI: implementation as learning. J Responsible Innov 6:375–380. https://doi.org/10.1080/23299460.2019.1603570

Emery S, Kendall H, Frewer LJ (2014) Independent evaluation of the public engagement with research and research engagement with society (PERARES) project

European Commission, Directorate-General for Research and Innovation, Delaney N et al (2020) Institutional changes towards responsible research and innovation: achievements in Horizon 2020 and recommendations on the way forward. Publications Office

European Commission (2014) Rome declaration on responsible research and innovation in Europe. https://digital-strategy.ec.europa.eu/en/library/rome-declaration-responsible-research-and-innovation-europe

Ferrari E (2020) Technocracy meets populism: the dominant technological imaginary of Silicon Valley. Commun, Cult Crit 13:121–124. https://doi.org/10.1093/ccc/tcz051

Franklin S (2015) Control: digitality as cultural logic. MIT Press

Fricker M (2007) Epistemic injustice: power and the ethics of knowing. Oxford University Press

Gastil J, Richards RC (2017) Embracing digital democracy: a call for building an online civic commons. PS: Polit Sci Polit 50:758–763. https://doi.org/10.1017/S1049096517000555

Gray M, Samala R, Liu Q et al (2024) Measurement and mitigation of bias in artificial intelligence: a narrative literature review for regulatory science. Clin Pharmacol Ther 115:687–697. https://doi.org/10.1002/cpt.3117

Hadar I, Hasson T, Ayalon O et al (2018) Privacy by designers: software developers' privacy mindset. Empir Softw Eng 23:259–289. https://doi.org/10.1007/s10664-017-9517-1

Hansson K, Ekenberg L, Cars G, Danielson M (2013) An e-participatory map over process methods in urban planning. https://libris.kb.se/bib/22275604
Hartz-Karp J, Briand MK (2009) Institutionalizing deliberative democracy. J Public Aff 9:125–141. https://doi.org/10.1002/pa.320
Held D (2006) Models of democracy. Polity Press, Boston/Medford, MA and New York, NY
Hofmann J (2019) Mediated democracy—linking digital technology to political agency. Internet Policy Rev 8:1–18. https://doi.org/10.14763/2019.2.1416
Hullman J, Diakopoulos N (2011) Visualization rhetoric: framing effects in narrative visualization. IEEE Trans Visual Comput Graph 17:2231–2240. https://doi.org/10.1109/TVCG.2011.255
Iandoli L, Quinto I, Spada P et al (2018) Supporting argumentation in online political debate: evidence from an experiment of collective deliberation. New Media Soc 20:1320–1341. https://doi.org/10.1177/1461444817691509
Janssen M, Brous P, Estevez E et al (2020) Data governance: organizing data for trustworthy artificial intelligence. Gov Inf Q 37:101493. https://doi.org/10.1016/j.giq.2020.101493
Kaminski ME (2023) Regulating the risks of AI. Boston Univ Law Rev 103:1347
Karpowitz CF, Raphael C (2016) Ideals of inclusion in deliberation. J Deliberative Democr 12. https://doi.org/10.16997/jdd.255
Kennedy R, Sokhey AE, Abernathy C et al (2021) Demographics and (equal?) Voice: assessing participation in online deliberative sessions. Polit Stud 69:66–88. https://doi.org/10.1177/0032321719890805
Kennedy H, Hill RL, Allen W, Kirk A (2016) Engaging with (big) data visualizations: factors that affect engagement and resulting new definitions of effectiveness. First Monday 21. https://doi.org/10.5210/fm.v21i11.6389
Köbis N, Starke C, Rahwan I (2022) The promise and perils of using artificial intelligence to fight corruption. Nat Mach Intell 4:418–424. https://doi.org/10.1038/s42256-022-00489-1
Kordzadeh N, Ghasemaghaei M (2022) Algorithmic bias: review, synthesis, and future research directions. Eur J Inf Syst 31:388–409. https://doi.org/10.1080/0960085X.2021.1927212
Kruse LM, Norris DR, Flinchum JR (2018) Social media as a public sphere? Politics on social media. Sociol Q 59:62–84. https://doi.org/10.1080/00380253.2017.1383143
Kučera E, Haffner O, Drahoš P et al (2020) New software tool for modelling and control of discrete-event and hybrid systems using petri nets. Comput Inform 39:568–586. https://doi.org/10.31577/cai_2020_3_568
Maciel C, Bicharra Garcia AC (2007) Design and metrics of a 'Democratic Citizenship Community' in support of deliberative decision-making. In: Wimmer MA, Scholl J, Grönlund Å (eds) Electronic government. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 388–400
Mazzucato M, Eaves D, Vasconcellos B (2024) Digital public infrastructure and public value: what is 'public' about DPI? https://www.ucl.ac.uk/bartlett/publications/2024/mar/digital-public-infrastructure-and-public-value-what-public-about-dpi
Miller D (2003) Deliberative democracy and social choice. In: Debating deliberative democracy, pp 182–199. https://doi.org/10.1002/9780470690734.ch9
Narechania TN, Sitaraman G (2024) An antimonopoly approach to governing artificial intelligence. Yale Law Policy Rev 43:95
Overton SA (2025) Analyzing the benefits of artificial intelligence to racially inclusive democracy. https://scholarship.law.gwu.edu/faculty_publications/1784
Owen R, Macnaghten P, Stilgoe J (2012) Responsible research and innovation: from science in society to science for society, with society. Sci Public Policy 39:751–760. https://doi.org/10.1093/scipol/scs093
Papadopoulos Y, Warin P (2007) Are innovative, participatory and deliberative procedures in policy making democratic and effective? Eur J Polit Res 46:445–472. https://doi.org/10.1111/j.1475-6765.2007.00696.x
Parthasarathy A, Phalnikar A, Jauhar A et al (2024) Participatory approaches in AI development and governance: a principled approach. arXiv.Org, June 3. https://arxiv.org/abs/2407.13100v1

Pedrini S, Bächtiger A, Steenbergen MR (2013) Deliberative inclusion of minorities: patterns of reciprocity among linguistic groups in Switzerland. Eur Polit Sci Rev 5:483–512. https://doi.org/10.1017/S1755773912000239
Perez O (2020) Collaborative e-rulemaking, democratic bots, and the future of digital democracy. Digit Gov: Res Pract 1. https://doi.org/10.1145/3352463
Petropoulos A, Pataki B, Juijn D et al (2025) Building cern for AI: an institutional blueprint. Centre for Future Generations, Brussels, Belgium. https://cfg.eu/building-cern-for-ai/
Quick KS, Feldman MS (2011) Distinguishing participation and inclusion. J Plan Educ Res 31:272–290. https://doi.org/10.1177/0739456X11410979
Rip A (2014) The past and future of RRI. In: Life sciences, society and policy, vol 10, p 17. https://doi.org/10.1186/s40504-014-0017-4
Risse M (2023) Political theory of the digital age: where artificial intelligence might take us. Cambridge University Press, Cambridge
Sanders LM (1997) Against deliberation. Polit Theory 25:347–376. https://doi.org/10.1177/0090591797025003002
Sastry G, Heim L, Belfield H et al (2024) Computing power and the governance of artificial intelligence. Sastry, Girish, Lennart Heim, Haydn Belfield, et al. 2024. arXiv.Org, February 13. https://arxiv.org/abs/2402.08797v1
Schmidt J (2024) The dignity circle: towards a non-ideal approach of social dignity in the context of marginalisation and care. PhD thesis. https://research.uvh.nl/en/publications/the-dignity-circle-towards-a-non-ideal-approach-of-social-dignity/
Shevlane T, Farquhar S, Garfinkel B et al (2023) Model evaluation for extreme risks. arXiv.Org, May 24. https://arxiv.org/abs/2305.15324v2
Shin D, Chotiyaputta V, Zaid B (2022) The effects of cultural dimensions on algorithmic news: how do cultural value orientations affect how people perceive algorithms? Comput Hum Behav 126:107007. https://doi.org/10.1016/j.chb.2021.107007
Shukla AK, Tripathi S (2024) AI-generated misinformation in the election year 2024: measures of European Union. Front Polit Sci 6
Silver H, Scott A, Kazepov Y (2010) Participation in urban contention and deliberation. Int J Urban Reg Res 34:453–477. https://doi.org/10.1111/j.1468-2427.2010.00963.x
Stilgoe J, Owen R, Macnaghten P (2013) Developing a framework for responsible innovation. Res Policy 42:1568–1580. https://doi.org/10.1016/j.respol.2013.05.008
Sujana IG, Lestari LD, Bara YP, Nagel Y (2025) Artificial intelligence and the future of democratic participation: opportunities and challenges. In: Proceeding of international conference on social sciences and education, pp 41–45
Vargas A, Lo A, Howes M, Rohde N (2017) The problem of inclusion in deliberative environmental valuation. Environ Values 26:157–176
Velikanov C (2017) Can deliberative governance become inclusive? In: Proceedings of the 18th annual international conference on digital government research. Association for Computing Machinery, New York, NY, USA, pp 531–540
von Schomberg R (2011) Towards responsible research and innovation in the information and communication technologies and security technologies fields. Publ Off Eur Union. https://data.europa.eu/doi/10.2777/58723
von Schomberg R (2013) A vision of responsible research and innovation. In: Responsible innovation: managing the responsible emergence of science and innovation in society. John Wiley & Sons, Ltd, New York, NY, pp 51–74
Wang G, Gold M, Siegel J et al (2015) Deliberation: obtaining informed input from a diverse public. J Health Care Poor Underserved 26:223–242
Weckert J, Valdes HR, Soltanzadeh S (2016) A problem with societal desirability as a component of responsible research and innovation: the "If we don't somebody else will" argument. NanoEthics 10:215–225. https://doi.org/10.1007/s11569-016-0258-1
Wilson C (2022) Public engagement and AI: a values analysis of national strategies. Gov Inf Q 39:101652. https://doi.org/10.1016/j.giq.2021.101652
Young IM (2002) Inclusion and democracy. OUP Oxford