Characterizing Democratic Biases in AI-Powered Participatory Democracy Platforms

Manon Berriche, Salim Hafid, Andreï Mogoutov, Jean-Philippe Cointet

Characterizing Democratic Biases in AI-Powered Participatory Democracy Platforms

Manon Berriche, Salim Hafid, Andreï Mogoutov, Jean-Philippe Cointet

Table of content

1. Defining the notion of "democratic biases" 3

1.1. Critical review of the literature on AI bias and the specificities of LLMs 4

1.1.1. Bias in Artificial Intelligence 4

1.1.2. Specificities of biases in LLMs 5

1.1.3. Limits of the term "bias" and of studies on AI/LLM bias 7

1.2. Conceptual and methodological implications 9

1.2.1. Distinguishing different types of bias 10

1.2.2. Distinguishing different levels of bias provenance 12

1.2.3. Adopting a hybrid, multi-layered, and continuous approach 13

1.3. LLMs and participatory-democracy platforms 14

1.3.1. LLM biases applied to political and democratic processes 15

1.3.2. Framing of the "AI for Democracy: Democratic Commons" project 17

1.3.3. Provisional definitions of the notion of "democratic biases" 18

1.3.4. Illustrative example based on existing literature 20

2. Identifying and characterizing democratic biases 22

2.1. Panorama of participatory-democracy platforms: investigating how democratic principles are translated into metrics and technical operations 22

2.2. Social choice theory: a set of robust tools for measuring adherence of AI systems to normative democratic desiderata 25

2.3. AI systems put to the test of citizens' critiques 28

2.4. The latent value matrix: Uncovering biases in how AI systems infer citizens' preferences 30

2.5. Modeling debate dynamics 32

3. Empirical studies 33

3.1 Algorithmic Approaches to Opinion Selection for Online Deliberation: A Comparative Study 33

3.2 Legibot: A comprehensive theory-driven annotated corpus and SLM for French legislative data 34

Bibliography 34

1. Defining the notion of “democratic biases”

The “biases” of artificial intelligence systems have sparked numerous controversies that have made headlines in recent years: think of sexist biases in automatic translation tools or racist biases in decision-support systems used in courts. The ubiquity of the term “bias” in the media sphere has contributed to keeping the concept conceptually vague. Yet the scientific literature has considerably refined the description and understanding of AI bias: for more than ten years, many studies have revealed diverse mechanisms, proposed operational typologies, developed appropriate metrics, and tested audit protocols to measure the systematic discriminations these systems can generate.

At a time when AI systems are becoming more complex in both design and deployment, it is urgent to rely on this body of work to specify the notion of “democratic biases” that we use in the “*AI for Democracy: Democratic Commons”* project. The aim is to propose an operational definition grounded in the conceptual reflections and empirical work on algorithmic bias, and to confront this definition with the application cases encountered on the participatory-democracy platforms we study. This conceptual work seeks to situate the notion within the current state of knowledge, to identify what the literature shows and does not show, and then to draw clear conceptual and methodological consequences for future empirical work.

The goal of this first section is to build that foundation. It is organized into three complementary parts. The first (1.1) offers a critical review of the literature on AI bias, emphasizing the specificities introduced by large language models (LLMs) and the conceptual and empirical limits of the term “bias”. The second (1.2) draws out the conceptual and methodological implications of this literature review: we distinguish different types of relevant biases, detail their multiple origins (design, data, model, interface, uses), and motivate the adoption of a hybrid, multi-layered approach to evaluation. The third part (1.3) confronts these findings with the field of participatory-democracy platforms: it synthesizes work specifically addressing the use of LLMs in participatory and deliberative processes, recalls the framing of the “*AI for Democracy: Democratic Commons* project”, and leads to an initial proposal for a modular definition of “democratic bias.”

The ambition of this work is both practical and reflective: to provide a common framework for the teams that is precise enough to guide comparable empirical measurements, while remaining adaptable to different disciplines, fields, and usages. The pages that follow deploy this logic by alternating theoretical synthesis, critical diagnostics, and methodological directions.

1.1. Critical review of the literature on AI bias and the specificities of LLMs

1.1.1. Bias in Artificial Intelligence

Over the past decade, the question of bias in artificial intelligence systems has given rise to a substantial body of work, mainly in the areas of fairness in machine learning and algorithmic justice. This research demonstrates that algorithmic systems are not neutral: they tend to reproduce, and sometimes amplify, preexisting social inequalities.

Foundational contributions, such as Barocas and Selbst (2016), show how training data can introduce indirect discrimination, notably in hiring or credit decisions. Angwin et al. (2016)'s study of the COMPAS tool revealed that recidivism scores systematically overestimated risk for Black individuals and underestimated it for white individuals. In computer vision, Buolamwini and Gebru (2018) found that facial recognition systems exhibited significantly higher error rates for women and darker-skinned people than for white men. Complementarily, Obermeyer et al. (2019) describe a different but equally problematic mechanism: a public-health algorithm used healthcare costs as a proxy for "medical need." Because structural factors cause some populations — here, Black patients — to receive less care on average and therefore incur lower costs, the algorithm systematically underestimated their true need for care. This example shows that bias does not only stem from demographic imbalance in data but can arise from an inappropriate choice of target variable (label/proxy bias), with concrete and potentially harmful consequences.

These findings have fueled debates among public commentators and scholars (in political science, philosophy, and computer science) about algorithmic fairness and the notion of *fairness* itself (Kleinberg et al., 2016; Beaudouin & Maxwell, 2023; Rabonato & Berton,

2025). Numerous criteria and metrics have been proposed (error-rate parity, equality of opportunity, demographic independence, etc.) to quantify, mitigate, or correct these biases (Mehrabi et al., 2021). However, these criteria can be mutually incompatible: choosing one notion of justice is inherently a political and normative decision that can render an algorithm "fair" by one metric and "unfair" by another. Theoretical results — some dating back several years — demonstrate that certain fairness definitions are formally irreconcilable, especially when the prevalence of a given trait is uneven across population subgroups (Kleinberg et al., 2016). Such impossibility results force trade-offs: one must prioritize certain criteria, make normative choices explicit, or seek technical and socio-political compromises.

Consequently, research on algorithmic bias yields two complementary lessons: first, the causes of bias are multiple (data, label choice, architecture, annotation processes, deployment modalities) and require nuanced diagnostics; second, technical measures and fixes are not value-neutral — any evaluation or intervention presupposes an explicit normative arbitration (which values to prioritize? which costs to accept?), which should be documented and deliberated with stakeholders prior to implementation.

1.1.2. Specificities of biases in LLMs

The emergence of large language models (LLMs) has shifted and complicated the research agenda on bias. Like other machine-learning models and systems, LLMs rely on training corpora whose composition, provenance and documentation can remain opaque; however, they differ in the scale and heterogeneity of those corpora, the complexity of their training pipelines, and their generalist nature — that is, their capacity to be applied to a wide diversity of tasks rather than to a narrowly defined objective (for example named-entity recognition or facial recognition). This combination — unprecedented data scale, documentary and procedural opacity, and generative inference mechanisms — makes identification, explanation and correction of biases far more difficult than with traditional classifiers designed for targeted tasks (Weidinger et al., 2021; Bommasani et al., 2021).

Several specific features explain this difficulty. First, the scale and heterogeneity of corpus composition (web crawls, forums, media, books, code, etc.) favor the inheritance and sometimes the amplification of historical stereotypes and under-representations — a

problem worsened by the practical impossibility of fine-grained curation at that scale. Moreover, structural data-collection practices introduce systematic skew: Bouchaud and Ramaciotti (2025) show that web-crawler restrictions (e.g., robots.txt entries, explicit scraping bans and rate limits) disproportionately exclude highly reliable, politically neutral sites from large corpora such as CommonCrawl, effectively increasing the relative share of more partisan or fringe sources in training datasets. Consequently, LLMs may both internalize dominant cultural norms and reproduce systematic omissions or amplified partisan signals (Gallegos et al., 2024; Bender et al., 2021).

Second, the generative and open nature of model outputs complicates bias identification: it is no longer sufficient to count mislabelled instances across well-defined socio-demographic categories. One must analyze the propensity of a generated text — typically in the form of a narrative — to convey stereotypes, erase minority voices, or present biased interpretive frames depending on prompts that are themselves open-ended, and on the conversation history each user maintains with the model. As shown by Bang et al. (2024), such biases cannot be captured solely by categorical error counts, since they operate both at the level of *what is said*—through selective emphasis, omission, or topic choice—and *how it is said*—through framing, tone, and rhetorical cues embedded in generated narratives. The high fluidity and persuasive force of generated content make such biases harder to detect, characterize or contest with simple tests. New forms therefore appear: hallucinations, discursive homogenization, and the over-representation of dominant languages and cultures in responses (Huang et al., 2025; Ferrara, 2023).

Third, beyond direct data transfer, alignment procedures — notably Reinforcement Learning from Human Feedback (RLHF) — add a poorly documented normative layer: the preferences and judgments of annotators and evaluators — often drawn from teams located in narrow geographic and socio-cultural contexts — guide the selection and ranking of outputs after training. This can produce a tendency to privilege majority opinions and systematically underrepresent minority perspectives in final outputs (Xiao et al., 2024; Santurkar et al., 2023; Kirk et al., 2024).

Finally, some observed phenomena are neither reducible to data nor to human preference but appear to stem from structural defects of the model itself. Boelaert et al. (2025), for

instance, describe “machine bias”: stable model behaviors — low variability of responses, biases that vary non-trivially with the question — that seem to originate in architecture, optimization, and internal generative mechanisms rather than as a simple reflection of social biases present in the training data or design choices. These machine biases can inject systematic noise into sensitive uses (e.g., simulated surveys, synthetic panels) and strongly complicate interpretation of diagnostics based solely on dataset composition. Relatedly, recent work suggests that safety and moderation layers can introduce their own structural distortions: Rogers and Zhang (2025) show that increased guardrail sensitivity induces a systematic “neutrality bias,” whereby ambiguous or contested political content is disproportionately classified as neutral, effectively flattening disagreement and masking conflict even in the absence of explicit partisan skew.

The specificities of LLMs make the measurement and quantification of bias more difficult than in classical machine-learning cases. Because LLMs combine massive, opaque corpora, plausible generative outputs, human alignment procedures, and emergent structural properties, standard metrics and common benchmarks are often insufficient or misleading. It is therefore essential not only to distinguish finely between types of bias and their sources, but also to devise new measurement frameworks and instruments to identify and characterize these biases.

1.1.3. Limits of the term “bias” and of studies on AI/LLM bias

A substantial body of work—primarily from STS, critical data studies, gender studies, and postcolonial studies—has challenged the standardized use of the term “bias.” Three major lines of critique emerge.

First, the term “bias” elides the political and situated nature of the problems: it suggests a technical, correctable error while obscuring design choices, power relations and social asymmetries, and sidestepping questions of purpose and accountability (Barocas & Selbst, 2016). Powles and Nissenbaum (2018) thus criticize the technical obsession with “solving” bias as an attractive diversion: by reducing deep political and economic questions (mass data capture, platform concentration, intended uses) to problems of metrics and optimization, we hide the power relations that produce these inequalities. They also show that

algorithmic “fixes” can be perverse—strengthening and refining instruments of surveillance and ranking—and advocate structural responses (data governance, transparency, external accountability mechanisms) rather than technical band-aids. This perspective is echoed by Barassi (2024), who argues that so-called “hallucinations” and large-scale AI failures should be understood not as isolated technical flaws, but as sociotechnical phenomena rooted in data practices, institutional incentives, and structural power relations.

Second, these critiques insist on the centrality of power and human labor: datasets, annotation categories, RLHF panels, and contracts with outsourced providers are all sites where perspectives are negotiated and inequalities reproduced (Miceli et al., 2022; Wirth et al., 2025).

Third, they point to the methodological fragility of dominant diagnostics: benchmarks and ad-hoc tests often yield results that are highly sensitive to prompt design, language, model version, and sampling parameters, undermining replicability and limiting their ability to predict real-world effects of LLM deployment (Blodgett et al., 2020; Barrie et al., 2024; Lum et al., 2024; Röttger et al., 2024).

Added to these critiques is a conceptual dimension emphasized by Gallegos et al. (2024): fairness approaches often rely on “implicitly desirable criteria”—for example, that outputs be independent of any social group—without making explicit the social and normative values that justify these choices. This shows that even technical evaluations reproduce implicit value choices, reinforcing the idea that the term “bias” conceals political and normative dimensions.

These critiques lead to several practical prescriptions for characterizing and auditing LLM biases: (i) make explicit the norm and objective adopted before any measurement; (ii) document the production chain (data, annotators, RLHF); (iii) favor contextualized and triangulated evaluations (metrics, discourse analysis, user studies); (iv) make visible and regulate the power relations and human labor that structure ongoing evaluation; and (v) situate the LLM outputs within their concrete interaction environment: through which interfaces they are delivered, via which material setups, and in which social or institutional situations they are mobilized.

These recommendations gain particular force in light of the literature on bias. A recent meta-analysis of 189 articles published between 2014 and 2024 (ACL, FAccT, NeurIPS, AAAI) shows that 82% of the studies do not formally define the "bias" they analyze, and that the majority (79.9%) focus on gender bias (often occupation–gender). This is followed by race/ethnicity (30.2%), age (20.6%), religion (19.1%), and nationality (13.2%) (Ghosh & Wilson, 2025). This concentration reflects a narrow conception of bias centered on a few highly institutionalized—and predominantly Western—social categories. Such focus tends to freeze identities into rigid boundaries (for example, a gender binary, homogeneous ethnic categories), limiting understanding of finer or intersectional forms of discrimination and rendering invisible situated experiences that escape these classifications (Dubet, 2016). In other words, the dominant "bias" approach risks reproducing, or even reinforcing, the classificatory frameworks it purports to critique. In this sense, research on bias in language models reproduces a methodological and conceptual bias of its own: it studies "biases" while neglecting the political, normative, and structural dimensions that produce them (Ducel et al., 2024).

Within the *"AI for Democracy: Democratic Commons"* project, it is important to take into account these critical contributions and their conceptual and methodological recommendations. Rather than abandoning the term "bias," we propose decomposing it into conceptual and operational families. This modularization has two practical advantages: (i) it makes explicit the assumptions and values behind each measure, facilitating transparency and deliberation about normative choices; and (ii) it guides the selection of appropriate audit protocols and remedies.

1.2. Conceptual and methodological implications

The literature devoted to bias shows that the issues are both plural—discrimination, invisibilization, amplification of dominant opinions, factual errors, etc.—and multilayered: they are rooted in data, expressed through algorithms, crystallized in design choices, and manifested differently depending on contexts of use. These findings call for a non-reductionist reading: biases cannot be confined to simple "technical errors" correctable by more data; they are complex, situated sociotechnical phenomena, produced and

reproduced at every stage of the design, training, and deployment of algorithms (Waseem et al., 2021; Weidinger et al., 2021).

Depending on the objective—diagnosing, measuring, correcting, or governing—it is often more relevant to adopt a differentiated vocabulary and appropriate conceptual instruments. As Lopez (2021) and Ferrara (2023) emphasize, grouping under a single label the effects of injustice, stereotyping, or discursive manipulation amounts to masking their distinct origins and heterogeneous normative implications. This is why several recent works (Mehrabi et al., 2021; Li et al., 2023; Chu et al., 2024; Guo et al., 2024) call for distinguishing both different types of bias and their levels of provenance. These distinctions make it possible to choose targeted metrics, audit protocols, and remedies rather than relying on generic approaches. This section lays out the conceptual clarifications and methodological choices needed to study biases in AI systems and LLMs: it proposes an operational typology of forms of bias (1.2.1), maps possible layers of origin (1.2.2), and defines a hybrid, multi-layered, continuous evaluation method (1.2.3).

1.2.1. Distinguishing different types of bias

Recent research on bias in AI and in LLMs offers many taxonomies, both for describing bias and for mitigating it (Chu et al., 2024; Gallegos et al., 2024; Bouchard, 2024). These classifications are not mutually exclusive: above all, they are characterization tools meant to guide operationalization—choice of metrics, evaluation protocols, correction strategies. As Lopez (2021) or Waseem et al. (2021) suggest, it is less fruitful to accumulate typologies than to organize them around analytical axes that link diagnosis and intervention.

Two axes prove particularly useful. The first concerns the affected groups and individuals (“who?”): sociodemographic biases (gender, race/ethnicity, age, religion, disability, socioeconomic status), linguistic and cultural biases (languages, dialects, registers), and political/ideological biases (implicit preferences, partisan framings). These dimensions often intersect: the political position of a group or individual can interact with their linguistic or socioeconomic status to produce specific effects in a public deliberation (Mehrabi et al., 2021). Several empirical investigations have highlighted the importance of these dimensions: Lucy & Bamman (2021) showed the reproduction of gender stereotypes in GPT-3, while

Rozado (2023, 2024), Hartmann et al. (2023), and Motoki et al. (2023) documented structural political orientations of language models that are often difficult to neutralize because social norms and regulatory pressure against political bias are weaker than for race or gender (Peters, 2022).

The second axis concerns modes of manifestation (“how?”). Three registers appear central. Representational biases refer to stereotyping, denigration, or erasure of certain groups (Caliskan et al., 2017; Bolukbasi et al., 2016). Allocative biases concern tangible distributive consequences, for example the visibility of contributions in a discussion thread or differential access to the agenda (Barocas & Selbst, 2016). Finally, discursive or epistemic biases affect the quality of information and deliberation through selective framing, omission, or the production of erroneous summaries (Weidinger et al., 2021). Again, these registers are far from independent and often intertwine: a representational omission can lead to a loss of access to the agenda (allocative) and impoverish the diversity of arguments available (discursive).

To operationalize these distinctions, it is essential to articulate two perspectives. The descriptive perspective consists in documenting observed phenomena—measuring distributions, quantifying performance or exposure gaps, identifying artifacts (for example, evaluating dialect coverage in automatic summaries). The normative perspective consists in determining, based on explicit values, whether and how to correct these gaps—choosing a justice criterion, defining an intervention objective, setting an acceptable threshold (Gallegos et al., 2024; Ghosh & Wilson, 2025). Without this articulation, bias measurement remains methodologically fragile and politically blind.

This double reading (descriptive diagnosis and normative judgment), crossed with the matrix “groups/individuals concerned × modes of manifestation”, provides a pragmatic tool for orienting the selection of metrics (disparity, coverage, worst-group accuracy, etc.), protocols (counterfactual tests, embedding probes, diversified panels), and interventions (data cleaning/enrichment, fairness techniques, interface adjustments, or governance mechanisms). In that sense, typologies are not meant to freeze airtight categories, but to equip a contextualized, multi-layer approach adapted to the diversity of LLM uses and social environments.

1.2.2. Distinguishing different levels of bias provenance

Beyond distinguishing different types of bias, it is crucial to map the layers where they appear. Suresh & Guttag (2021) and Hovy & Prabhumoye (2021) propose frameworks that identify sources of harm throughout the model life cycle; applied to LLMs, these frameworks make the multi-layered nature of the problems visible.

At the design level, normative and technical choices crystallize—task definitions, optimization objectives, selection of metrics and architectures. These decisions often embed implicit values and determine which forms of pluralism will be taken into account or neglected. As Rieder & Skop (2021) note, biases can thus emerge not only from code or data, but also from the norms and hypotheses adopted by design teams.

Choices made when defining learning objectives, inputs, outputs, and success criteria (the form of the chosen loss function) can introduce implicit biases, thereby steering the model toward potentially unfair or discriminatory outcomes.

Training data biases result from unequal historical and structural representations in the corpora used for training, such as Wikipedia or Common Crawl (Hovy & Prabhumoye, 2021; Blodgett et al., 2020). Some groups may be over or under-represented, leading to stereotyped associations and to models that reflect and amplify these imbalances.

Annotation and post-training processes, including RLHF, incorporate human preferences that can reinforce dominant norms and introduce new biases (Hovy & Prabhumoye, 2021; OpenAI, 2023).

At the model level, embedding biases endogenize stereotyped associations, for example between occupations and gender (Bolukbasi et al., 2016; Zhao et al., 2018). Biases can also appear in content generation or response ranking, directly affecting the model's outputs.

Interface design and interaction modes steer user attention and influence actual uses of the model, potentially creating biases in how system recommendations are perceived or adopted (Rieder & Skop, 2021).

Finally, biases can manifest in the real-world use of AI systems depending on the institutional, social, or media contexts into which they are integrated. For example, certain populations may be systematically marginalized or their contributions ignored in practical applications (Sap et al., 2019).

To clarify this complexity, Guo et al. (2024) and Li et al. (2023) propose distinguishing between intrinsic biases (embedded in the model's internal representations) and extrinsic biases (observable in downstream tasks and context-dependent). This distinction, though imperfect, helps guide interventions: intrinsic biases call for solutions during training or data curation, while extrinsic biases require remedies at the interface, post-processing, fine-tuning, or governance levels.

1.2.3. Adopting a hybrid, multi-layered, and continuous approach

The preceding analyses show that LLM biases come in varied types and emerge at different levels of provenance, from the choice of data and architectures to actual uses in social and institutional contexts. For instance, Feng, Park, Liu, and Tsvetkov (2023) trace how political biases travel from pretraining corpora through model representations to downstream tasks, demonstrating that provenance-aware audits and systematic dataset provenance tracking are essential for diagnosing and mitigating unfair political effects in NLP pipelines (Feng et al., 2023). This complexity makes insufficient any approach limited to a one-off or exclusively technical audit: understanding and addressing biases requires simultaneously articulating technical diagnostics, socio-organizational investigations, and normative considerations. In other words, evaluating large language models must be conceived as a continuous and layered activity. It involves not only measuring observable performance and gaps, but also integrating social and organizational perspectives so as to link observed phenomena to norms of justice and representativeness. This approach, inspired by recent prescriptions (Mökander et al., 2024; Wirth et al., 2025), aims to provide a framework capable of guiding targeted interventions at each stage of the LLM life cycle—from design and training to deployment and use—while ensuring coherence between descriptive diagnostics and normative judgments.

Two guiding ideas structure this approach. The first is simple: separate levels of analysis to better connect diagnostics to remedies. Mökander and coauthors (2024) propose a reading in three complementary layers—governance, model, application—which makes it possible to examine, for each system, (i) who decides and under which rules, (ii) what the model does and how it does it, and (iii) how it is concretely used in a given context. This grid invites us not to confuse the solution (e.g., a sampling adjustment) and responsibility (who validates and documents this choice). In that sense, the technical audit becomes one piece of a broader apparatus of institutional accountability. The second major idea is continuity: evaluation must be integrated into the system life cycle rather than treated as an isolated event. Wirth et al. (2025) show how much "algorithmic production" is a process of continuous evaluation, where datasets, alignment criteria, deployment parameters, and uses evolve and mutually re-evaluate one another. In practice, this implies setting up permanent monitoring, feedback loops, and formal update procedures—including alert thresholds, correction protocols, and reversibility mechanisms—to prevent drift from settling in between audits.

The *"AI for Democracy: Democratic Commons"* project, structured into teams from complementary disciplines, naturally lends itself to this multi-layered approach. Each team cannot on its own cover all bias types or all levels of origin, but, based on a shared conceptual template, each can make explicit what it is working on—type of bias targeted, preferred level of analysis, metrics used, evaluation methods, etc.

1.3. LLMs and participatory-democracy platforms

After clarifying the limits of the term "bias" and proposing a typological and multi-layered approach, it is now appropriate to place the reflection back into the specific context that concerns us: the growing use of LLMs in participatory and deliberative democracy processes. Introducing LLMs into these arenas is not neutral. It raises new questions: how do the biases documented in the previous sections manifest when the goal is to process political content or facilitate collective deliberation? To what extent do these biases affect representativeness, fairness, and the legitimacy of democratic processes? And conversely,

what potential do LLMs offer to broaden participation, foster inclusion, or improve deliberative quality, provided they are integrated into robust institutional architectures?

This section addresses these issues in four steps. First, we will examine the available empirical results concerning LLM biases when applied to political and democratic processes (1.3.1). Next, we will present the conceptual and methodological framing of the *"AI for Democracy: Democratic Commons"* project (1.3.2). Finally, we will propose a provisional definition of the notion of "democratic biases" (1.3.3), illustrated by a concrete example (1.3.4).

1.3.1. LLM biases applied to political and democratic processes

While artificial intelligence is being deployed across many domains of everyday life, recent years have seen a rapid proliferation of AI tools designed specifically to support democratic processes — for instance through summarization, moderation, translation, and rephrasing (for reviews see Konya et al., 2023; Goldberg et al., 2024; Aoki, 2024). Citizens are increasingly turning to LLMs during civic activities: voters may query a model about party platforms before casting a ballot, or use it to draft or polish political messages for social media.

Empirical work now shows that these interactions are not neutral. Controlled and survey-based studies indicate that LLM outputs can shape political attitudes and behaviors: Fisher et al. (2024) document that biased model outputs can meaningfully influence political decision-making and deliberative conduct, and Potter et al. (2024) provide causal evidence from experiments that subtle framing, rhetorical style, and contextual cues in model responses can produce measurable short-term shifts in voter preferences. Together, these findings argue that democratic-bias assessment must go beyond measuring model outputs and explicitly evaluate user-level outcomes.

At the same time, descriptive audits of model outputs paint a complex and sometimes contradictory picture of political orientation. Some studies report systematic leanings (e.g., left-leaning tendencies reported in Santurkar et al., 2023; Hartmann et al., 2023), while

others show that apparent bias depends strongly on model scale, prompt design, evaluation task, and the language or cultural framing used (Wright et al., 2024; Rettenberger et al., 2025). Large comparative investigations further demonstrate that political bias is not reducible to a single left–right axis: different model families and tasks exhibit uneven and task-dependent patterns, which supports the need for contextualized, multi-model, and reproducible audits rather than one-off partisan labels (Peng et al., 2024).

The risks are particularly salient for automated summarization and aggregation tasks used in participatory settings. Several studies show that LLMs can omit important contributions, over-generalize, or reframe content in ways that change meaning (Huang et al., 2024; Steen & Markert, 2023; Maynez et al., 2020). Other work demonstrates that even when factual coverage is adequate, framing effects — stylistic choices, rhetorical devices and pragmatic tone — can systematically shift interpretive valence and thereby affect downstream political judgments (Bang et al., 2024). Conversely, some sophisticated system designs and interaction protocols (e.g., iterative critique or deliberative validation) can help preserve a wider range of positions in generated summaries (Fish et al., 2023; Tessler et al., 2024), highlighting that sociotechnical arrangements shape outcomes as much as model internals do.

A related strand investigates whether LLMs can substitute for human respondents in social-science tasks such as surveys or simulations. Results are ambivalent: some experiments show that models conditioned on demographic profiles can reproduce aggregate response distributions (Argyle et al., 2023), while others conclude that models cannot reliably replace human subjects due to unstable biases across topics or because they better approximate subgroups with near-uniform statistics (Domínguez-Olmedo et al., 2024; Boelaert et al., 2025). Architectural choices and the presence (or absence) of demographic priors also matter — Park et al. (2024) report high fidelity in a specific agent-based setup, suggesting that representativeness is achievable but contingent on design and evaluation choices.

Taken together, these empirical and descriptive findings warrant three interlinked responses. First, audits must be task-specific, comparative, and reproducible, combining output diagnostics with user-level outcome measures (attitude change, vote intention, deliberative

behavior). Second, evaluation should explicitly include framing and stylistic metrics in addition to content coverage and factuality, because framing biases can produce political effects even when content-level metrics look satisfactory. Third, and importantly, the problem should be recognized as partly structural: Peters (2022) frames algorithmic political bias not merely as an empirical anomaly but as a feature emerging from design choices and governance arrangements, implying that remedies require institutional and normative interventions (transparency, procedural safeguards, and accountability mechanisms) rather than only technical patches.

In short, assessing and mitigating LLM democratic risks demands integrated sociotechnical protocols that link model audits to interaction design, human oversight, and governance frameworks — otherwise seemingly minor output differences can translate into substantive democratic effects.

1.3.2. Framing of the “AI for Democracy: Democratic Commons” project

Within the *“AI for Democracy: Democratic Commons”* project, initial framing elements proposed by team members laid the groundwork for a common analytical grid intended to orient and make empirical work comparable. This framing first identifies five operational missions that AI can assume in democratic environments—moderation, summarization, translation, writing assistance, and debate facilitation—and lists thirteen use cases requiring dedicated audits, spanning participation in the citizen agenda, participatory budgeting, and understanding and drafting legislation (Alspektor & Mas, 2025).

In parallel, the project articulates five democratic principles that serve as a normative compass: participation; political and ethical pluralisms; deliberation; responsibility (capacity to act and accountability); and recognition of agreements and disagreements. These principles are supported by four cross-cutting requirements—transparency, agency, autonomy, and equality—which guide the choice of metrics, investigation protocols, and recommendations (Pénigaud & Reber, 2025).

Recalling this framing before defining the notion of “democratic bias” is essential: it shifts the exercise from the theoretical register to the operational register. The definition we propose must be usable by each project team, meaning it must be translatable into measurable indicators, audit protocols, and mitigation scenarios relevant to a given use case.

Finally, this conceptual architecture imposes a strong methodological rule: every evaluation must be contextualized. One does not evaluate an “LLM” abstractly, but rather evaluates a specific role of AI (e.g., summarization or moderation), deployed in a specific use case, and in light of a given democratic principle. This requirement of contextualization conditions the choice of baselines, metrics, and recommended interventions, and it ensures that diagnoses of “democratic bias” will be rigorous, comparable, and actionable.

1.3.3. Provisional definitions of the notion of “democratic biases”

Rather than imposing a single definition of “democratic bias,” we propose below a set of modular definitions: a shared theoretical definition serving as a normative reference, followed by operational definitions intended to be adopted and adapted by each team depending on its discipline, fieldwork, and methodological objectives. The goal is to provide both a shared framework and concrete instruments—measurable and interpretable—allowing comparable diagnoses, while respecting epistemologies and local investigation contexts.

A democratic bias can be defined as a measurable deviation from a chosen baseline (e.g., empirical, normative, technical, regulatory), produced by a sociotechnical assemblage which, in a given use case, for an engaged user or public, compromises, distorts, or hierarchizes—intentionally or not—the respect of a targeted democratic principle.

This formulation emphasizes three essential dimensions. First, contextuality: a democratic bias is not an intrinsic property of an abstract model but the result of an interaction between a functionality (e.g., summarization, moderation), a concrete use case, and the sociotechnical conditions of deployment. Second, normativity: calling an output “biased” implies explicitly referring to democratic principles and an ethical justification—one must say

which norms are being protected and why. Third, multicausality: observed deviations can emerge from combinations of technical causes (architecture, optimization), organizational causes (RLHF panels, moderation policies), and social causes (power relations, inequalities of access).

Operationally, any diagnosis of democratic bias must therefore (a) make the targeted principle explicit, (b) describe the AI's role and the use case studied, (c) make explicit and justify the chosen baselines, and (d) specify the metrics used to measure LLM outputs relative to the chosen reference norm (descriptive representativeness vs. normative correction). Once this diagnosis is established, one can think about a remediation solution (e) by identifying plausible mechanisms and the relevant levels of analysis.

To make this definition more concrete, we can illustrate the reasoning with an example of a summarization task whose purpose is to respect the principle of pluralism. During a citizen consultation on ecological transition, an LLM is used to produce a summary intended for public decision-makers and the general public. Examination of this summary reveals that contributions favorable to economic regulation measures (for example introducing a carbon tax or restricting road traffic) are developed and illustrated with concrete examples, while critical contributions (concerns about purchasing power, defense of individual freedom) are condensed into a single generic sentence, without nuance.

Different choices of baselines make it possible to evaluate this gap:

- **Empirical baseline:** the real distribution of contributions shows that about 35% of texts express critical positions. The fact that they occupy less than 10% of the summary constitutes an obvious descriptive deviation.
- **Normative baseline:** the organizers specified in the protocol that "all significant positions must be represented in the final summary, regardless of their frequency." The observed deviation then becomes a democratic bias, since it explicitly violates the principle of pluralism.
- **Technical baseline:** comparison with an earlier version of the summarization system shows that it reproduced about 25% of critical arguments, versus less than 10% with

the current version. The bias thus appears as a technical regression that degrades argumentative coverage.

- **Regulatory baseline:** certain national or local rules governing consultations require that minority opinions be made visible in public reports. By not respecting this obligation, the summary becomes not only democratically biased but also legally non-compliant.

The analysis highlights a multicausal origin: the model's tendency to privilege expert-style argued formulations, prompt settings that favor seeking consensus rather than making disagreements visible, and the absence of a specific review aimed at checking viewpoint diversity. Remediation avenues include adding explicit instructions in the prompt to guarantee representation of all identified positions, developing argumentative-coverage indicators measuring the relative presence of each opinion category, and implementing targeted human validation focused on respect for pluralism before releasing the summary.

1.3.4. Illustrative example based on existing literature

We previously discussed how democratic biases can affect the functioning of deliberation and citizen-participation platforms. In this section, we illustrate, through a concrete example, how such biases can emerge. We structure this section around three axes: (i) a task that can be automated or assisted by LLMs, (ii) a sociotechnical context in which the LLM is deployed and where citizens interact with the LLM and/or with each other, (iii) a democratic principle that risks being violated.

We again choose task (i): summarizing citizens' opinions. This task can be formulated as follows: given a corpus of opinions in the form of free-text written by human participants, the LLM must produce one or more summaries that are representative of the distribution of opinions in the initial corpus.

In the existing literature, this task has been operationalized in various sociotechnical contexts (ii): for example, in the "Habermas Machine" (HM) experiment (Tessler et al., 2024), a virtual citizens' assembly was constituted using a demographically representative sample

of the United Kingdom. The HM, a trained (fine-tuned) LLM, acts as a "caucus mediator": citizens interact with the LLM but never directly with each other. More precisely, citizen–LLM interactions are structured as follows: citizens individually express their opinions in free text; the LLM generates a summary based on all citizens' opinions; then citizens individually critique the LLM-generated "group summary." Citizens' critiques are then fed back to the LLM, which produces new, improved group summaries. In separate research, Fish et al. (2024) formalize the problem of automated opinion summarization by linking it to properties of democratic representation. They mathematically demonstrate verification of these properties under the condition of an "oracle," i.e., an LLM capable of perfectly modeling every citizen's individual preferences. In practice, they conducted pilot experiments based on surveys: citizens provided opinions in free text, and an LLM produced a series of representative summaries. As in the HM case, citizens do not directly interact with each other, and evaluation of LLM-generated summaries is done via validation surveys rather than synchronous face-to-face deliberation. Unlike the HM protocol, aggregation and validation of generated summaries are not iterative and do not incorporate citizens' critiques. Revel & Pénigaud (2025) complement these empirical approaches with a normative analysis: they argue that the democratic legitimacy of LLM-assisted summarization depends on making the system's underlying design choices intelligible and open to public assessment (and, where appropriate, consent), and they caution against deployments that treat AI-produced outputs as binding or that risk disempowerment and post-rationalization. Together, these studies show how the same technical task—summarizing human opinions—can be integrated into sociotechnical contexts that offer very different interaction patterns (direct iterative critique vs. survey + validation; private caucus vs. public deliberative debate) and varying degrees of citizen autonomy.

Moreover, experiments from existing work highlight the main democratic biases that can emerge from using LLMs in such sociotechnical contexts. For example, consider pluralism as the democratic principle (iii). Does the summary generated by the LLM preserve the diversity of viewpoints, including minority, dissenting, or unconventional perspectives? Potential violations of this principle can be envisioned due to various phenomena. Considering only phenomena linked to the LLM as a technical object—i.e., a large pretrained language model—violations could arise from undue compression of expressed opinions,

where a size-limited summary risks erasing nuance and making minority opinions invisible in favor of the majority opinion alone. Paraphrasing could also play a role, since reformulating citizens' opinions might induce tone shifts that result in an ideological drift.

By focusing only on the task of opinion summarization by LLMs, this example shows that while LLMs can help quickly surface a consensus and efficiently reorganize diverse opinions, the democratic legitimacy of their results depends strongly on concrete sociotechnical choices—nature of interactions, sampling, weighting, provenance, contestability, and human oversight. In the absence of explicit safeguards tied to democratic criteria (for example, proportional participation rules, pluralism guarantees, channels for iterative critique), LLMs are not exempt from generating illusory consensus or eradicating minority voices. Only an ad hoc and situated examination of these risks—attentive to potential deviations documented in the literature as well as to still-unknown effects—can guard against them and establish genuine trust between citizens and these participation and deliberation platforms.

Bibliography

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., ... & McKee, K. R. (2024, May). The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias — There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica.

Aoki, G. (2024). Large Language Models in Politics and Democracy: A Comprehensive Survey. arXiv preprint arXiv:2412.04498.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.

Ashkinaze, J., Fry, E., Edara, N., Gilbert, E., & Budak, C. (2025, April). Plurals: A system for guiding llms via simulated social ensembles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-21).

Alspektor & Mas. (2025). *[Cadrage projet AI for Democracy: Democratic Commons]*. [Rapport interne / working paper].

Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*.

Barassi, V. (2024). Toward a theory of AI errors: Making sense of hallucinations, catastrophic failures, and the fallacy of generative AI. *Harvard Data Science Review*, (Special Issue 5).

Barocas, Solon, & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104(3),* 671-732.

Barrie, C., Palmer, A., & Spirling, A. (2024). Replication for language models problems, principles, and best practice for political science.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of the ACM FAccT Conference

Beaudouin, V., & Maxwell, W. (2023). La prédiction du risque en justice pénale aux États-Unis: l'affaire ProPublica-Compas. *Réseaux*, *240*(4), 71-109.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint* arXiv:2005.14050.

Boelaert, J., Coavoux, S., Ollion, E., Petev, I., & Präg, P. (2025). Machine Bias. How Do Generative Language Models Answer Opinion Polls?. *Sociological Methods & Research*, 00491241251330582.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NeurIPS* (2016).

Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint* (2021).

Bouchaud, P., & Ramaciotti, P. (2025). Web Crawler Restrictions, AI Training Datasets\& Political Biases. arXiv preprint arXiv:2510.09031.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.

Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter*, 26(1), 34–48.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *Journal of Machine Learning Research*, *24*(312), 1-117.

De Domínguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2024). Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37, 45850–45878.

Ducel, F., Névéol, A., & Fort, K. (2024). La recherche sur les biais dans les modèles de langue est biaisée: état de l'art en abyme. *Revue TAL: traitement automatique des langues*, *64*(3).

Dubet, F. (2016). *Ce qui nous unit: discriminations, égalité et reconnaissance.* Seuil.

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283*.

Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Fish, S., Gö lz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., & Wüthrich, M. (2023). Generative Social Choice. *arXiv preprint* arXiv:2309.01291.

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., ... & Reinecke, K. (2024). Biased ai can influence political decision-making. arXiv preprint arXiv:2410.06415.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097-1179.

Ghosh, S., & Wilson, K. (2025). Bias is a Math Problem, AI Bias is a Technical Problem: 10-year Literature Review of AI/LLM Bias Research Reveals Narrow [Gender-Centric] Conceptions of'Bias', and Academia-Industry Gap. *arXiv preprint arXiv:2508.11067*.

Goldberg, B., Acosta-Navas, D., Bakker, M., et al. (2024). AI and the Future of Digital Public Squares. *arXiv preprint* arXiv:2412.09988.

Gudiño, J. F., Grandi, U., & Hidalgo, C. (2024). Large language models (LLMs) as agents for augmented democracy. *Philosophical Transactions A*, *382*(2285), 20240100.

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., ... & Xiong, D. (2023). Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2020. World Values Survey: Round Seven – Country-Pooled Datafile. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi.org/10.14281/18241.1

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., … Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55.

Hovy, D., & Prabhumoye, S. (2021). *Five sources of bias in natural language processing. Language and Linguistics Compass, 15 (8), Article e12432*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (ITCS'17).

Konya, A., Turan, D., Ovadya, A., Qui, L., Masood, D., Devine, F., ... & Forum, D. A. (2023). Deliberative Technology for Alignment. *arXiv preprint arXiv:2312.03893*.

Li, Y., Du, M., Song, R., Wang, X., & Wang, Y. (2023). A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Liu, R., Jia, C., Wei, J., Xu, G., & Vosoughi, S. (2022). Quantifying and alleviating political bias in language models. *Artificial Intelligence*, *304*, 103654.

Lopez, P. (2021). Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, *10*(4), 1-29.

Lucy, L., & Bamman, D. (2021). *Gender and representation bias in GPT-3 generated stories*. In *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48–55). Association for Computational Linguistics

Lum, K., Anthis, J. R., Robinson, K., Nagpal, C., & D'Amour, A. (2024). Bias in language models: Beyond trick tests and toward ruted evaluation. *arXiv preprint arXiv:2402.12649*.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). *On faithfulness and factuality in abstractive summarization*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919)

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*.

Miceli, M., Posada, J., & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power?. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1-14.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: a three-layered approach. *AI and Ethics*, *4*(4), 1085-1115.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. *Public Choice*, *198*(1), 3-23.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., … Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint* arXiv:2411.10109.

Peng, T. Q., Yang, K., Lee, S., Li, H., Chu, Y., Lin, Y., & Liu, H. (2024). Beyond Partisan Leaning: A Comparative Analysis of Political Bias in Large Language Models. arXiv preprint arXiv:2412.16746.

Pénigaud, T., & Reber, B. (2025). *[Document projet Democratic Commons — principes et exigences]*. (rapport interne / preprint).

Peters, U. (2022). Algorithmic political bias in artificial intelligence systems. *Philosophy & Technology*, *35*(2), 25.

Poole-Dayan, E. G. (2025). From Dialogue to Decision: An LLM-Powered Framework for Analyzing Collective Idea Evolution and Voting Dynamics in Deliberative Assemblies. *Master's thesis, MIT.*

Potter, Y., Lai, S., Kim, J., Evans, J., & Song, D. (2024). Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. *arXiv preprint arXiv:2410.24190*.

Powles, J., & Nissenbaum, H. (2018). The seductive diversion of 'solving' bias in artificial intelligence. *Onezero*.

Rabonato, R. T., & Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics*, 5(3), 1943–1959.

Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. Journal of Computational Social Science, 8(2), 1-17.

Revel, M., & Pénigaud, T. (2025). AI-Facilitated Collective Judgements. *arXiv preprint* arXiv:2503.05830.

Rieder, B., & Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society*, *8*(2), 20539517211046181.

Rogers, R., & Zhang, X. (2025). A bias towards neutrality? How LLM guardrail sensitivity affects classification. Communication and Change, 1(1), 13.

Röttger, P., Hinck, M., Hofmann, V., Hackenburg, K., Pyatkin, V., Brahman, F., & Hovy, D. (2025). IssueBench: Millions of realistic prompts for measuring issue bias in LLM writing assistance. *arXiv preprint* arXiv:2502.08395.

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, *12*(3), 148.

Rozado, D. (2024). The political preferences of LLMs. *PloS one*, *19*(7), e0306621.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023, July). Whose opinions do language models reflect? In *International Conference on Machine Learning* (pp. 29971–30004). PMLR.

Sharma, T., Potter, Y., Kilhoffer, Z., Huang, Y., Song, D., & Wang, Y. (2024). From experts to the public: Governing multimodal language models in politically sensitive video analysis. *arXiv preprint* arXiv:2410.01817.

Small, C. T., Vendrov, I., Durmus, E., Homaei, H., Barry, E., Cornebise, J., … Megill, C. (2023). Opportunities and risks of LLMs for scalable deliberation with Polis. *arXiv preprint* arXiv:2306.11932.

Steen, J., & Markert, K. (2024, August). Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 5962-5983).

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the ML lifecycle. *Proceedings / arXiv* (2021).

Taubenfeld, A., Dover, Y., Reichart, R., & Goldstein, A. (2024). Systematic biases in LLM simulations of debates. *arXiv preprint arXiv:2402.04049*.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., … Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, 386(6719), eadq2852.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Waseem, Z., Lulz, S., Bingel, J., & Augenstein, I. (2021). Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*.

Wirth, A., Ludec, C.L., Girard-Chanudet, C. *et al.* Who Decides When Artificial Intelligence Works? Algorithmic Production as a Continuous Evaluation Process. *Digital Society.* 4, 65 (2025).

Wright, D., Arora, A., Borenstein, N., Yadav, S., Belongie, S., & Augenstein, I. (2024). LLM Tropes: Revealing fine-grained values and opinions in large language models. *Findings of ACL: EMNLP 2024*, pp. 17085–17112.

Xiao, J., Li, Z., Xie, X., Getzen, E., Fang, C., Long, Q., & Su, W. J. (2024). On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *arXiv preprint* arXiv:2405.16455.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *NAACL-HLT* (2018).