KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

AI and Democratic Publics

Henry Farrell and Hahrie Han

August 2025

Sébastien A. Krier using Midjourney 6.1

Introduction

Could existing democratic institutions and processes be improved by AI? A burgeoning body of scholarship asks how AI-driven machine learning can improve—or even replace—democratic institutions that aggregate opinions and beliefs (Ovadya 2023, Jungherr 2023).

This literature makes strong, but often unstated assumptions about how democracy works, and where it can go wrong, creating a tacit paradigm that guides scholars to focus on some questions, problems, and hypotheses at the expense of others. As one of us has argued

together with co-authors in the past:

> Paradigms guide action. Particularly in moments of crisis, those paradigms—or cohered sets of assumptions about ourselves, each other, and the world around us—shape the intentions we develop, the solutions we imagine, and, ultimately, the actions we choose. What happens when the paradigms we carry are limited or, worse, wrong? … [Paradigms] illuminate possibilities for change, they also constrain where we look. The wrong paradigm leads us to misread situations, overlook opportunities, and pursue the wrong solutions. (Vallone et al 2023)

In this paper, we argue that the existing paradigm of democracy driving scholarship about its relationship to AI highlights the wrong questions. The essay describes this broad paradigm—which emphasizes the benefits of deliberation and sortition—and explains why it is insufficient for understanding or acting in a healthy democracy. We argue that we should instead focus on enduring democratic publics and how they shape collective behavior. That would raise very different questions. How might AI reshape these publics and the feedback loops that they depend on? Will this contribute to democratic stability or undermine it? Such questions would underpin a broader and different research agenda on AI and democracy than the one we have today.

The Dominant Paradigm and Its Limitations

Much prominent research on AI and democracy emphasizes two lines of inquiry: sortition and deliberation. Sortition is the random selection of a group of citizens representative of the broader population, while deliberation is reasoned debate among citizens on matters of common concern. Scholars such as Hélène Landemore (2020) have argued that sortition and deliberation can be combined to generate "mini-publics"—citizen assemblies that can

come together for a limited period to deliberate on important public questions and reach democratically legitimate conclusions. These ideas have inspired AI 'labs' to fund research on the relationship between AI and democracy, which they define in terms of deliberation and sortition (OpenAI 2023). Landemore (2024) has furthermore speculated that AI could be deployed to facilitate deliberation at scale, so that large percentages of the population could participate in rolling mini-publics, enabling a kind of democracy in which citizens better govern themselves.

This work has important insights about one aspect of democracy—deliberation—but also notable limitations as a complete model of the complex processes and practices that constitute a healthy democracy. These limits become clear in the inability of deliberative-sortition approaches to deliver in practice. It is telling that the most notable effort to date to use sortition to address democratic discontent and disaffection—Macron's efforts to institute deliberative assemblies in France—appears to have failed (Raymond 2022).

Deliberative-sortition approaches are right to focus on publics, but they miss out on important aspects of how actual publics work, and how democracy reproduces itself. At their core, deliberative-sortition approaches are built on models of disinterested individuals making reasoned arguments. This renders collective formations necessary only as a form of temporary aggregation.

It is no accident that deliberation-sortition has emerged as a starting point for the discussion about AI and democracy, because its underlying model of politics comports easily with dominant technocratic approaches to addressing (well-founded) contemporary alarm about polarization. Although 'reason' (in the sense used by deliberation scholars) and 'rationality' (in the sense used by technocrats) are not, actually, the same thing, they often

blur together. The engineers developing AI favor technocratic approaches that emphasize efficient aggregation of preformed beliefs, and the way people's capacity to reason allows them to talk together and reach some optimal understanding of the public good (Austen-Smith et al, 1996). Both similarly stress the value of many different points of view in solving complex problems, whether this involves debate or market-based means for discovering the 'wisdom of crowds.'

Likewise, the promise of deliberation offers a neat solution to the problem of 'epistemic polarization,' where citizens radically disagree over what is true. Deliberation-sortition advocates argue that mini-publics and similar institutions are superior ways to synthesize knowledge. By recruiting a sufficiently large and diverse body of citizens at random, mini-publics can represent the public at large, and take advantage of its diversity of opinions and knowledge to reach better solutions, while avoiding the kinds of polarization that regular politics entails. Thus, some scholars argue that "open democracy" is inherently superior to other democratic forms in identifying scalable solutions to problems in a complex world (Landemore 2020, 8). In the language of computer science, mini-publics are 'lossy representations' of the public as a whole: living photographs of the public that blur out some of the details but capture the main features of the larger and far more complex entity that they stand in for. AI may make such representations more accurate by efficiently scaling up the publics, correcting common epistemic pathologies to guide diverse and polarized publics toward common agreement on politically contentious questions, and using dialogue to obviate conspiracy theories (Landemore 2024; Tessler et al. 2024; Costello, Pennycook, and Rand 2024).

Mini-publics appeal to the undoubtedly attractive but implausible notion that we can scrub power relations from democracy. More bluntly still: they suggest that we can have

democracy without politics (as politics is ordinarily understood).

Both deliberation theorists and technocrats tend to have idealized understandings of the cognitive, emotional, and motivational capacities that citizens can contribute to democratic politics. Under both accounts, cool reasoning replaces the jostle of politics. They idealize the notion of a rational, well-educated citizen, who has sufficient individual competences to reason about her own interests, the public interest, and the relationship between the two. Individuals' private capacities for reason then, can lead to collectively beneficial outcomes either because people reason together for the collective good, 'leaving their interests at the door,' or because some mechanism of aggregating individual beliefs leads to collectively superior outcomes. That is why these accounts are so appealing: it would be lovely to replace the messy complexities of politics and power amongst humans with something more reasonable.

Although large language models (LLMs) and other forms of AI could help aggregate and summarize people's beliefs so as to discover optimal points of agreement, deliberative approaches tend to assume away the agonism of politics (the basic fact that it involves different organized factions competing for power). They miss the underlying implicit and explicit power dynamics that shape the construction, expression, and impact of such opinions. To be stable and effective, both democracy and democratic publics require people to be socially embedded in *ongoing* relationships rather than temporary mini-publics. Such relationships may be less efficient in the short term, so that some optimal policies remain out of reach, but they are better over time at generating political compromises that can weather adversity. Democracy is not just challenged to solve problems better but to ensure its own stability.

We emphasize that this does not exclude the important empirical research that is being done around LLMs. To pick one prominent example, the "Habermas Machine" created by Tessler et al. 2024 is something genuinely new in the world, modeling an innovative and potentially important form of representation that can be applied across many issues. However, we argue this research agenda will be more fruitful and useful if it does not limit its attention to deliberation-sortitionist accounts, and instead engages actively with the variety of forms of actually existing democracy, the overlapping but different aspirations that we might have for democratic institutions, and the different ways in which they function. In particular, we emphasize the importance of enduring democratic publics based around shared coalitions and reconciliation of clashing interests, rather than the more evanescent publics centered on abstracted discussion that deliberation-sortitionists emphasize. Understanding how these new technologies are taken up, and how they may be made more useful will require a broader understanding of democracy in practice.

The deliberation-sortition approach depicts democracy as legitimate primarily insofar as it encourages reasonably disinterested conversation—when randomly selected bodies of citizens can deliberate together to find their way to common understandings of the problems that they face through exercising voice, or to allow them to vote with their feet or otherwise, exiting arrangements they dislike. Political forms such as political parties, social movements, and civic associations are regularly dismissed as outmoded and inefficient forms of indirect representation that can be swept away as soon as we have better technologies of aggregation, consensus building, and representation.

As Albert Hirschman argues in his canonical book, *Exit, Voice, and Loyalty* (1970), enduring collective interests are enormously important. Political forms such as political parties, social movements, and civic associations function by generating not disinterest but

loyalty. Democratic politics requires citizens to develop shared commitments so that they can negotiate across the different perspectives that are bound up in the contest for voice and power, both within and between groups. It is precisely because people recognize that they are involved in a repeated game that they become willing to negotiate solutions to their differences. That is why enduring collective forms such as social movements and political parties have persisted despite repeated efforts to replace them with more efficient arrangements. These vehicles of collective action are not simply technical apparatuses for deliberating about and aggregating individual beliefs and opinions—instead, they are the structures through which citizens form, negotiate, and cultivate both their own and others' citizenship. Publics are not more-or-less accurate momentary snapshots of citizens' opinions and knowledge, but processes which evolve over time, and can work out well, or badly, depending.

Hence, we argue that any model of democracy should embrace the politics involved in managing power instead of avoiding them. Accepting the everyday social, collective dimensions of political behavior and disputes over power allow us to understand politics better than models that try to assume away innate human tendencies. This—as we now argue—points toward a very different way of thinking about how AI may reshape the world.

An Alternative Paradigm of Politics

To clarify our argument, we draw a stronger contrast between the dominant deliberation-sortitionist approach and our own than their actual differences warrant. Even so, they have distinct central tendencies and questions. Our own guiding assumption is that democracy is a socially embedded phenomenon in which people develop their preferences, understandings of politics, and meaning through collective organization around interests and shared beliefs. People are not individualistic, disinterested actors, but instead motivated by

a desire to realize their own interests in an ongoing collective competition for power. This points to a different understanding of the problems to be solved in democracy. If we take the fact of political competition seriously, we understand that democracy is a game played into the indefinite future. Without the possibility of repeated cooperation, individual and collective actors could easily defect from the game when it does not go their way, possibly leading to its (democracy's) breakdown. If I, as a leader of one major political faction, do not think that you, as the leader of another will give up power if you lose an election, then I may be unwilling to give up power myself. Losers in a competitive game must have incentives to keep playing. In democracy, the incentive is the ongoing ability to negotiate their own individual interests within enduring processes of competition, collaboration, and compromise that hold out some reasonable prospect of prevailing in the future (Przeworski 1991). This means that episodic deliberation can plausibly supplement competition and collaboration under some circumstances, but it cannot replace it. Because deliberative approaches ignore political competition and sideline individual interests, they are poorly suited to figure out how to strike bargains that balance competition against compromise.

More generally, people do not participate as individuals in the political process, but as members of groups (Farrell, Mercier and Schwartzberg 2023). People do not typically arrive at their preferences and political beliefs through introspection. Nor are people necessarily devoted to democratic politics as such (Elliott 2023). Instead, we argue, their preferences are socially constructed through engagement with others in their families and social groups. These and broader groups are necessary not only to formulate interests and positions, but to cultivate ongoing commitments to a variable political process that inevitably involves negotiation and compromise away from the initially formulated positions (Ahlquist and Levi 2013). Likewise, people's approach to politics is usually more practical. When they get engaged in politics, it is not because they want to debate the best things to do from first

principles, but because they have attached themselves to a group or a cause (Han 2024). Putting the pieces together, our approach sees humans as social beings inevitably enmeshed in collective processes that both shape and help realize their interests. If the particular kinds of deliberation-sortition accounts favored by many AI researchers start from rational citizens willing to deliberate in selfish or selfless ways about how to achieve their own individual interests and the common good, our account assumes that in what political scientist Robert Dahl (1963) calls "the great circus of life," the cultivation and protection of democracy risks becoming nothing but a mere "sideshow." Making democracy work, then, is less about cultivating rational or reasonable individuals with a set of capacities to act in particular ways, but instead about the constitution and interaction of particular problem-focused publics—or collectives that bring people together around shared problems and become vehicles through which their interests are constructed, expressed, and transformed.

As pragmatists like John Dewey (2012) and Jane Addams (1902) argued, publics are living, ever-evolving groups, rather than the briefly animated representative snapshots of broader opinion that mini-publics provide. Building on Dewey, we define a public as an evolving, living relationship between the people who collectively constitute it, and the feedback loops (technologically mediated or otherwise) that allow them to understand themselves as a collective and act accordingly. Dewey sees publics as a response to problems stemming from the "objective fact that human acts have consequences upon others, that some of these consequences are perceived, and that their perception leads to subsequent effort to control action so as to secure some consequences and avoid others" (Dewey 2012, 66). A potential public consists of all those who are affected by a particular social problem and have some practical interest in addressing it. A public comes into being when people recognize that the problem exists, and act together to resolve it. And that, in turn, is the mainspring of

democracy. In Dewey's description:

> Full education comes only when there is a responsible share on the part of each person, in proportion to capacity, in shaping the aims and policies of the social groups to which he belongs. ... [D]emocracy ... is but a name for the fact that human nature is developed only when its elements take part in directing things which are common, things for the sake of which men and women form groups—families, industrial companies, governments, churches, scientific associations and so on. (Dewey 1948, 209)

Such publics create political feedback loops between what people do and how they think. Feedback loops thus provide crucial mechanisms through which a set of individuals become a collective. They educate people by creating spaces through which people rehearse ways of acting collectively to address shared goals. Perhaps most important are the feedback loops that exist between the beliefs, goals, and passions of the individual members and the collective understandings through which they orient themselves toward the world. As people work together, they come to roughly shared understandings of their common problems and common desires not through deliberative processes that require them to leave their particular interests at the door, but instead by learning how to negotiate their own interests within the context of competition, collaboration, and compromise.

Treating these groups as the relevant unit of analysis means that the core problems to be solved in democracy involve working with and through action-oriented publics that are formed around interpretation, passion, and shared commitment. Building such publics involves practical engagement with the relevant communities, as demonstrated by Jane Addams, who deeply influenced Dewey's philosophy of democracy; wrote widely in her

own right; and did more than Dewey ever could to figure out the practicalities of how to engage with communities and help women, immigrants, and other groups to organize in pursuit of shared goals. The publics that people like Addams helped bring into being were not the product of randomized selection, but particular groups with specific interests, which also had capacity to organize and ally with other groups, and contend with other groups than those. As Addams (1902, 172) remarked:

> In the disorder and confusion sometimes incident to growth and progress, the community may be unable to see anything but the unlovely struggle itself.

Two practical examples—one involving organizing around an algorithm, one involving organizing in the broader public space—illustrate what such publics might involve. In his book, *Voices in the Code*, David G. Robinson (2022), provides an ethnography of the public that developed around a proposed algorithm to allocate kidneys to transplant patients across the U.S.[1] This was a highly contentious algorithm: the weights that it gave to different factors (age, medical history, and similar) might determine who lived and who died. Robinson describes how the shift to algorithmic decision-making leads, all too often, to a watering-down of democratic control. However, as he stresses (8), "the rules and values for how we live together are a topic for a wider conversation … that … *cannot* … belong only to people who are fluent in technological terms of art."

As Robinson explains, debates over the kidney transplant algorithm provide an example of how an informed and *interested* public can form around a shared problem. This public was

1 We are grateful to Nathan Mathias for bringing this work to our attention.

born in heated contention among people with different values and interests. The debate over the algorithm was initially dominated by doctors and scientists, who found themselves having to share space with patients and donor families, all of whom had different understandings of the problem and how to solve it. Arriving at an acceptable compromise between different perspectives with different understandings of equity took a decade, far longer than anyone anticipated. Nor was everyone entirely happy with the final result. Nonetheless the community that formed around the algorithm arrived at a rough consensus that most of those involved could accept.

In Robinson's (101) description, the "debate both reflected and shaped the community's beliefs about what an acceptable algorithm would look like. The competing ideals … are ultimately 'irreconcilable . . . at the end, we make some sort of sausage.'" Notably, this community was an *interested* one, both because those involved had a strong interest in the outcome, and because their interests differed over what that outcome would be. Figuring out how to reconcile those interests and act on them was what made the community into a public. Furthermore, it was a community of long duration, in which the different groups were repeatedly thrown together, so that they developed some sense of common purpose and understanding, despite their disagreements. That shared sense was what allowed them to be efficacious.

A second example of how actual publics work today comes from Minnesota. ISAIAH is a statewide, multi-faith organizing group that was part of a broader effort to pass the 2023 "Minnesota Miracle" (Dionne 2023), a suite of legislation that secured one of the strongest social safety nets in the country. ISAIAH was a core part of a statewide multiracial coalition that came together—and stayed together despite attacks designed to divide it—to pass these policies despite having only a one-vote majority in the state Senate. ISAIAH itself consists of

a number of distinct sub-constituencies, which are each organized around a distinct institutional base. In a recent interview, former Executive Director Doran Schrantz described it as follows:

> Within ISAIAH, we have the Muslim coalition, which is 46 Islamic centers and neighborhood and community organizations that are Muslim, predominantly East African. There is the Rural Organizing Project, which is organizing working-class white rural people in key places around the state. There's Kids Count On Us, which is 500 community-based child-care centers organizing child-care providers, workers, and parents — kind of a union-type, proto labor formation for childcare in the state of Minnesota. We have the church base, with church clusters all over the state and anchored in the suburbs and the cities. We have the young adult coalition, which is 10 college campus configurations along with young adult apartment organizing. It's become an intergenerational, multiracial, multifaith, cross-geographical coalition. What holds ISAIAH together is a set of principles that we all agree to. Number one is the rigorous focus on...a north star of building multiracial, democratic governing power. (Schrantz, Doran, and Han 2024)

Each of those sub-constituencies represents its own Deweyian public, but they all come together to form a larger public within ISAIAH. Yet, given the diversity of all the constituencies, they inevitably have distinct interests—race, faith, age, and geography are just a few of the possible lines of division.

Instead of trying to paper over or assume away those distinctions, ISAIAH built its coalition by creating space to constantly negotiate those interests. Schrantz notes that this is not

just a "list of organizations" as many coalitions are, but instead is built on the "political work" of "holding collectives and constituencies together" with a shared focus on power (Schrantz, Doran, and Han 2024). Put another way, ISAIAH intentionally created processes to nurture the feedback loops that would constitute the public itself.

The repeated iteration of feedback loops is crucial because the goal is, as organizers commonly put it, not to get people (or publics) to "do a thing" but instead to "become the kind of person [or public] who does what is to be done" (Woodly 2022). Unlike mini-publics that bring people together to "do a thing" (deliberate) once, these Deweyan publics equip people and publics with the kind of internal compass they need to navigate uncertain, dynamic political environments in an ongoing way. Questions of self-interest and collective interest, of political power, and of the construction, maintenance, and expression of collective voice are central to this work. As Schrantz describes it:

> Inside ISAIAH … all of the organizers and lead organizers have a space to meet together for four hours every other Monday. It's an intensive space — it's not *Here's your task, and here's your task*. It's practical skills training. It's also political formation in relationship to each other — the Muslim coalition, the Black barbershop coalition, the white churches. We are negotiating questions like *What are our politics? What is our base? How do we operate with one another?* Within that space, there's a whole set of practices that are rehearsed and rehearsed and rehearsed.
>
> We also have a central leadership team of representatives from each constituency base that meets once a month for three hours. That is a political space in which their strategic orientation is cultivated. In that space, leaders have the

> experience of negotiation, working through actual problems, resolving differences, asking power questions. *What would this mean for you? How are we regulating our orientations with each other so that we viscerally experience our sense of shared responsibility face-to-face?* It's not abstract. It's literally people. It's a space of constant practice where people's own self-interest is bridged into the collective interest. We're constantly working to construct a collective common interest. (Schrantz, Doran, and Han 2024)

Without these feedback loops, Schrantz argues, the coalition would have been susceptible to breaking apart at the first sign of division. In the contemporary era of polarization and division, external political pressures could easily cause particular constituencies to "reduce back to silos by identity, faith, geography, etc." (Schrantz, Doran, and Han 2024).

Indeed, in 2022, when the Supreme Court overturned *Roe v. Wade*, ISAIAH faced a significant challenge. They were starting to build toward the statewide campaign that would result in the suite of legislation that became the 2023 Minnesota Miracle, but the *Dobbs* decision threatened to tear apart the base that would make such a political majority possible. Within ISAIAH, the progressive Christians were incensed by the Supreme Court's decision, but the Muslim coalition, parts of ISAIAH's rural base, and the Black barbershops were supportive of it. As Executive Director, Schrantz had to bring leaders from all of these constituencies together to decide how they were going to react in this moment. Would they let their divisions pull their coalition apart, or would it hold together? Schrantz describes the challenge:

> In that moment, we were building on the years of work we had done of constantly rehearsing people's understanding of how their self-interest is tied to a

> collective interest. How does that play out when it's hard? In this moment, these suburban white women felt they had an interest that is somehow being thwarted by this coalition: *I am part of this thing, and now it is constraining me in acting on my actual interest* … Can we actually have mutual interest in this moment? (Schrantz, Doran, and Han 2024)

In the end, through a set of complex discussions, the group decided to stay together by engaging in a process of rearticulating their shared goals and negotiating the places where they express their own interests (Schrantz, Doran, and Han 2024). Crucial to these discussions was a strategic understanding of power and collective interest that had been built through years of those shared meetings in which leaders developed visceral understandings of their own individual and collective power and learned to constantly negotiate individual and shared interests. Repeated practice built a set of muscles that enabled ISAIAH to hold its coalition together when tested.

Understanding publics in this way points toward broader forms of negotiation—of interests, identities, policies, self-understandings, commitments, and so on—than deliberative accounts imply. People's interests are formed not exogenously but endogenously within the process, but they are not expected to check those interests at the door. Instead, they negotiate new forms of collective identity in which they reshape their individual understandings of their interests with respect to the collective, and vice versa. Again: they are not disinterested—they may be passionate in their attachments to a group that they may come to see as an extended "community of fate," making common cause with members (Ahlquist and Levi 2013).

As demonstrated by ISAIAH, publics formed around such practically oriented forms

of democratic activity are likely to be more robust and long-lasting than deliberative mini-publics, since they are conjoined by shared interests rather than random sortition, or the kinds of aggregation and summarization of opinion that sortition-focused AI offers. Their partiality is both a potential weakness (they do not see the whole) and a source of enormous strength (they have more concrete appeal to ordinary citizens). They can—when they work—generate a self-sustaining internal political economy, vying to solve the problems of their members. They also may work together where there is common purpose: in Tocqueville's description, "In democratic countries, the science of association is the mother science; the progress of all the others depends on the progress of that one" (1990, 110). As a general matter, publics are built through feedback loops—what Schrantz referred to as the work of "rehearsing" a way of interacting—not just among the citizens that they bring together, but between those citizens and their leaders; the shared understandings, interests, and passions that allow them to coordinate; and the individuals who do the coordinating (Schrantz, Doran, and Han 2024).

By taking political conflict, difference, and struggle seriously, this alternative approach seeks to create a model of politics that more closely mirrors natural human behavior. The sortitionist utopia is a mirage. Citizen juries and the like can play an important subsidiary role, but they cannot be a model for politics as a whole. We need, instead, to ask how to bring publics and their interactions to the center of analysis and action. This then leads us to ask how technologies such as AI affect the construction of publics. We suggest that they affect the ways in which the public represents itself to itself. That leads us to consider some basic issues of AI and politics in a different light. Specifically, we can think more systematically about how they act as technologies that affect the feedback loops through which publics organize themselves.

AI, Feedback Loops, and Democracy

Obviously, publics can go wrong as well as right. The experience that Schrantz describes within ISAIAH is not the modal experience of most constituency-based groups in contemporary politics. In fact, one plausible diagnosis of what is amiss with U.S. politics, and the politics of many other advanced democracies, is that there is something wrong with how publics have come to be understood and constructed in contemporary political systems. The opportunities our system provides for ordinary citizens to come together to collectively solve problems, and the kind of groups that have emerged, have become hollowed out. Furthermore, the relationships *between* these different publics, with their clashing diagnoses of what the problems are, have become disordered.

People's ordinary experience of politics in America today is not of publics, but spectacle. Within a larger attention economy, such a political system treats citizens as consumers of political spectacle that feeds off outrage, treats identities as two-dimensional, and focuses on expression rather than efficacious action. When political action becomes purely expressive, it ignores the possibility of a complex set of feedback loops that can dynamically connect people to collectives that give them the political power to realize their individual and collective interests. As consumers, people act collectively only when the market aggregates their outrage into something additive. When those interests diverge, as ISAIAH's did after the *Dobbs* decision, market-like aggregations dissolve. Lacking durability, the aggregations are thus necessarily more fragile, and less powerful as vehicles of political expression. Without other options to exercise power, people are left only with the choice to express their outrage more vehemently. Transcending such spectacle requires citizens to rise heroically above their own interests.

Indeed, history teaches us that such publics can be crucibles of either democracy or authoritarianism. Civil society entities in the Weimar Republic, for instance, enabled the rise of Nazism in enormously destructive ways (Berman 2006). We need, then to understand the conditions under which publics are more likely to nurture democratic behavior and when they are more likely to dissolve or even break away from the democratic game. Comparing the experiences of people in an organization like ISAIAH to the experiences of people who are merely consumers of political spectacle points to the importance of the feedback loops that shape publics.

It would be worth developing a more explicit theory of feedback loops and representation, but this is not the paper to do that. Suffice to say that organized democratic publics (or, for that matter, other kinds of organized factions) require some feedback loop between the common beliefs that bind people together and the individual beliefs that their members and particular factions hold. The democratic system within which publics contend for power has its own shared beliefs, which again will differ from the beliefs of particular democratic publics. A healthy democracy requires sufficient sense of common purpose both within and across democratic publics that inevitable clashes and disagreements can be mediated without destabilizing the whole. Efficacious publics—like ISAIAH and the community that formed around the kidney transplant algorithm—require sufficiently strong feedback loops that people feel that their positions and desires have been sufficiently respected and considered. A public that degenerates into clashing factions instead of resolving disagreements will not be effective in pursuing its members' goals. Similarly, a democratic system that degenerates into hostile clashes between different publics will at best be incapable of getting things done. At worst, it may collapse entirely.

This, then, allows us to begin to ask a different set of questions about the relationship

between AI and democracy. The feedback loops through which publics form and contend with each other are typically mediated through technology. As Perrin and McFarland (2011, 89) emphasize, "[p]ublics are evoked, even shaped, by [the] techniques that represent them." The cohesion of publics has depended, for example, "on technical practices such as the salons and coffee shops in Habermas …, electoral rituals …, media representations …, and standardized representative surveys." DeDeo (2017, 8) argues:

> Once we realize that the machine-aided predictors of a system are also participants, it is natural to ask how their use of that knowledge, accurate or not, back-reacts on the society itself. … We understand very little about how the introduction of these prediction algorithms, on a large scale, will lead to novel feedbacks that affect our political and social worlds; it remains an understudied and entirely open topic.

We can meld these arguments to ask how AI affects the kinds of feedback loops through which democratic publics are formed and interact with each other. In other words: how does AI affect the *internal constitution* of publics, and how does it affect *relations between* publics in the broader democratic system. To understand this, we want to focus on how the feedback loops within and between publics are increasingly mediated through AI and related technologies, which stand between the individuals and groups that constitute democracy.

Focusing on the consequences of AI for the internal constitution of publics might lead us to examine how AI and other algorithms are used to substitute for traditional bonds that hold coalitions together. For example, AI is being applied further to enhance microtargeting strategies, which seek to optimize votes, helping campaigns get enough votes to

win elections by sending customized appeals to particular subgroups of the population (Simchon et al. 2024). Such strategies scale more easily and cheaply than community building approaches. Although this might be more effective in turning out votes in the short-term (Enos, Fowler, Vavreck 2014), it plausibly creates aggregations of voters that are fragmented, based on market segmentation rather than common alliances and authentic or enduring social interactions. Such distorted publics are almost by definition internally disconnected, and likely to be more brittle in the face of challenges that affect subgroups differently, since they have little sense of solidarity or common purpose to call on.

Optimal AI strategies for particular campaigns such as microtargeting may thus weaken the collective capacities that enable publics to mediate and represent the interests of their members. More subtly, short term mini-publics might do the same, prioritizing temporary and specific inquiries over the longer-term commitments and loyalties that allow democracies to handle unexpected problems. Longer-term citizen assemblies would plausibly do better—but they would do so by creating the kinds of enduring collective interests that deliberation-sortitionist approaches undervalue or sometimes even actively disparage. Equally, it is possible that other applications of AI might strengthen rather than weaken group cohesiveness. Might the "Habermas Machine" (Tessler et al. 2024) help identify areas of common purpose across different groups in the decidedly non-Habermasian setting of an interested public? It is, at the least, possible. Might AI be more useful in solving problems if it is visibly biased in favor of one perspective than if it seeks common ground (Lai et al. 2025)?

Such questions are sorely underexplored as are their practical implications for technology and society, because they fit poorly within the existing paradigm of AI democracy. More generally, we have remarkably little sense of the new organizational forms that emerge as

AI applications are combined with traditional organizational tools (Farrell et al. 2025). The 'labs' and companies developing AI are obsessed with scaling, like the rest of the technology industry (Bracy 2025), perhaps suggesting that these technologies will exacerbate existing problems of disconnection and non-representativeness (Skocpol 2003, Karpf 2012, Skocpol, Williamson and Coggin, 2011; although see also Hall 2022).

Similarly, there is intense speculation about whether algorithms play an important role in political polarization in the United States and elsewhere. These discussions are disconnected from debates over AI and democracy, even though they largely concern the application of AI. The AI recommendation algorithms that social media services such as YouTube rely on, look, among other goals, to increase 'engagement' with content along various measures. This has led to speculation that they direct people down 'rabbit holes' toward toxic or extreme material that is both highly engaging and politically radicalizing (Lewis 2024). On the one hand, public debate emphasizes the likelihood of AI-fueled polarization, based in part on leaked discussions within social media companies which suggest that they have seen evidence of radicalization (Hao 2021). However, a data driven literature suggests that exposure to radical content is driven by demand rather than algorithms, so AI and related algorithms have a limited effect on public opinion (Budak et. al 2024).

Such putative algorithmic radicalization might indeed help explain why feedback loops between publics have become more problematic. It might increase these publics' unwillingness to maintain a shared democratic system, as they stop believing that they have interests in common. Equally, it is possible that different forms of AI such as LLMs could operate differently. The effects that Costello, Pennycook and Rand (2024) observe in an experiment where AI substantially lowers people's beliefs in conspiracy theories might be replicated on a larger scale if LLMs tend to guide people toward more mainstream beliefs on the basis

of some combination of post-training processing (reinforcement learning with human feedback) and innate characteristics of the models themselves (which tend to emphasize common aspects of the cultural data that they are trained on). Equally, as Resnik (2024, 3 ; see also Farrell 2025) puts it, "a lot of what's in people's heads sucks...and, crucially... LLMs have no way to distinguish the stuff that sucks from the stuff that doesn't." Neither reinforcement learning nor training provides any guarantees of outcome (training varies for models such as Elon Musk's Grok while common occurrence is no guarantee of rightness or non-bias). Incidents such as Grok's announcement that it was "MechaHitler" after an updated training prompt provide stark evidence of how badly LLMs can go wrong. Investigating the consequences of such effects for democratic publics is a research agenda rather than a set of predetermined conclusions.

Concluding Thoughts

To understand the consequences of AI for democracy, we need to move away from an existing debate that focuses heavily—and arguably nearly exclusively—on sortition, deliberation and disinterested mini-publics, to one that emphasizes the highly interested publics that drive politics in actually existing democracies. There is a strong observable bias in existing debates about democracy and AI against traditional political parties and social movements and the like and in favor of new forms of aggregation.

This leads to a variety of blind spots. As others have suggested (Revel and Pénigaud 2025), there is remarkably little understanding of the complications that lurk behind such apparently straightforward terms as "collective judgment" and "collective will." There is no developed literature on the benefits and costs of AI for traditional forms of democratic organization. Instead, there is an emphasis on moving away from such forms as quickly as possible toward better and more efficient ways of doing things. We need to stop focusing

exclusively on how AI might enable putative new forms of public that allow us to escape the vexations of politics, and start thinking about, and carrying out research on, how AI is feeding back into existing ones.

This would allow us to understand how AI may reshape democratic publics, in practice rather than in theory, by mediating traditional feedback loops in different ways or actively creating new ones. We could address existing debates about why democracies are in trouble by examining the role of technologies such as AI in exacerbating or mitigating the problems. We could think more systematically about how publics that are malign or deranged occur.

Finally, we could bring together engineers and social scientists in new intellectual configurations. Engineers often have a limited appreciation of how irreducibly messy politics are. This sometimes leads them to enthuse about visions of democracy—such as deliberation—that promise to eliminate the mess. Social scientists, including political scientists, are typically far more interested in studying problems than in trying to figure out how to solve them. We would be much better off if we could create frameworks in which engineers and computer scientists would focus their energies on mitigating the problems of democracy rather than hoping to sweep them away, while social scientists engaged in discovering solutions rather than empirical regularities. Understanding the consequences of actual AI for actual democracy, and figuring out how to steer toward the better rather than the worse outcomes, will require new kinds of cooperation than the ones we have today.

All this is for starters: we present these ideas in the hope that others, as well as criticizing and correcting errors, might point to other implications and applications. And we warmly acknowledge that much existing research (again: see Tessler et al. 2024) that emphasizes sortition may also generate insights about group phenomena in general that could be

extremely useful to the kinds of democracy that we believe to be essential. But to first see the problems that we identify, we need to be aware of the limitations of an influential paradigm that tends to limit inquiry into AI and democracy to one narrow question: how to apply AI techniques to make sortition and deliberation better.

Bibliography

Addams, Jane (1902), *Democracy and Social Ethics* (Macmillan).

Ahlquist, John S., and Margaret Levi (2013). *In the Interest of Others: Organizations and Social Activism* (Princeton University Press).

Austen-Smith, David, and Jeffrey S. Banks. "Information Aggregation, Rationality, and the Condorcet Jury Theorem." *American Political Science Review* 90, no. 1 (1996): 34-45.

Berman, Sheri (2006). *The Primacy of Politics: Social Democracy and the Making of Europe's Twentieth Century* (Cambridge University Press).

Budak, Ceren, Brendan Nyhan, David M. Rothschild, Emily Thorson and Duncan Watts (2024), "Misunderstanding the Harms of Online Misinformation," *Nature* 630:45-53.

Bracy, Catherine (2025). *World Eaters: How Venture Capitalism is Cannibalizing the Economy*. Penguin.

Costello, Thomas H., Gordon Pennycook, and David G. Rand (2024). "Durably Reducing Conspiracy Beliefs Through Dialogues With AI." *Science* 385.6714 (2024): eadq1814.

Dahl, Robert (1963). *Modern Political Analysis* (Prentice-Hall).

de Tocqueville, Alexis (1990). Democracy in America. New York: Vintage Books, 1990.

DeDeo, Simon (2017). "Major Transitions in Political Order." *From Matter to Life: Information and Causality*: 393.

Dewey, John (2012). *The Public and Its Problems: An Essay in Political Inquiry*. Penn State Press, 2012.

Dionne, E.J. (2023), "The 'Minnesota Miracle' Should Serve as a Model for Democrats," *Washington Post*, June 4, 2023.

Elliott, Kevin J. *Democracy for Busy People*. University of Chicago Press, 2023.

Enos, Ryan D., Fowler, Anthony, and Vavreck, Lynn. 2014. 'Increasing Inequality: The Effect of GOTV Mobilization on the Composition of the Electorate'. *Journal of Politics* 76(1):273–288

Farrell, Henry. "AI as Governance." *Annual Review of Political Science* 28, no. 1 (2025): 375-392.

Farrell, Henry, Alison Gopnik, Cosma Shalizi and James Evans (2025), "Large AI Models are Cultural and Social Technologies," *Science* 387:6739: 1153-1156.

Farrell, Henry, Hugo Mercier and Melissa Schwartzberg (2023), "Analytic Democratic Theory: A Microfoundational Approach," *American Political Science Review* 117,2 :767-772.

Hall, Nina (2022). *Transnational Advocacy in the Digital Era: Think Global, Act Local*. Oxford University Press, 2022.

Han, Hahrie (2024). *Undivided: The Quest for Racial Solidarity in an American Church*. Knopf: New York, 2024.

Hao, Karen (2021). "How Facebook Got Addicted to Spreading Misinformation," *MIT Technology Review*, https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

Hirschman, Albert O. Exit, *Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, 1970.

Jungherr, Andreas. "Artificial Intelligence and

Democracy: A Conceptual Framework." *Social Media+Society* 9.3 (2023): 20563051231186353.

Karpf, David (2012). *The MoveOn Effect: The Unexpected Transformation of American Political Advocacy*. Oxford University Press.

Lai, Shiyang, Junsol Kim, Nadav Junievsky, Yujin Potter, and James Evans (2025), *Biased AI Enhances Human Decision-Making But Reduces Trust* (unpublished paper).

Lewis, Becca. "Rabbit Hole: Creating the Concept of Algorithmic Radicalization." In Digital Media Metaphors, pp. 90-102. Routledge, 2024.

Landemore, Hélène (2020). Open Democracy: Reinventing Popular Rule for the Twenty-First Century (Princeton University Press).

Landemore, Hélène (2024), "Can Artificial Intelligence Bring Deliberation to the Masses," in Ruth Chang and Amia Srinivasan (ed.), *Conversations in Philosophy, Law, and Politics* (Oxford University Press).

Perrin, Andrew J., and Katherine McFarland (2011). "Social Theory and Public Opinion." *Annual Review of Sociology* 37.1:87-107.

Mercier, Hugo, and Dan Sperber (2017). *The Enigma of Reason*. Harvard University Press.

OpenAI (2023), "Democratic Inputs To AI," https://openai.com/index/democratic-inputs-to-ai/.

Ovadya, Aviv. "Reimagining democracy for AI." *Journal of Democracy* 34.4 (2023): 162-170.

Przeworski, Adam (1991). *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. Cambridge University Press.

Raymond, Gino Gérard. "Bottom-Up Democracy, Blame and a Republican Monarch Among the 'Déclassés'." *Modern & Contemporary France* 30, no. 4 (2022): 411-425.

Resnik, Philip. "Large Language Models Are Biased Because They Are Large Language Models." *Computational Linguistics Special Collection: Cognet* (2025): 1-21.

Robinson, David G (2022). *Voices in the Code: A Story about People, Their Values, and the Algorithm They Made*. Russell Sage Foundation.

Revel, Manon and Théophile Pénigaud (2025), *AI-Enhanced Deliberative Democracy and the Future of the Collective Will*, https://arxiv.org/pdf/2503.05830

Schrantz, Doran, and Hahrie Han (2024), "Our Power is Organized People," *Hammer and Hope* No. 5.

Simchon, Almog, Matthew Edwards, and Stephan Lewandowsky (2024). "The Persuasive Effects of Political Microtargeting in the Age of Generative Artificial Intelligence." *PNAS Nexus*, pgae035.

Skocpol, Theda (2003). *Diminished Democracy: From Membership to Management in American Civic Life*. Oklahoma University Press.

Tessler, Michael Henry, et al (2024). "AI Can Help Humans Find Common Ground in Democratic Deliberation." *Science* 386.6719: eadq2852.

Vallone, Dan, Hahrie Han, Emily Campbell, and Isak Tranvik. "Searching for a New Paradigm: Collective Settings." A report published by the SNF Agora Institute and More in Common. Baltimore, MD. 2023

Williamson, Vanessa, Theda Skocpol, and John Coggin. "The Tea Party and the Remaking of Republican Conservatism." *Perspectives on Politics* 9.1 (2011): 25-43

Woodly, Deva R (2022). *Reckoning: Black Lives Matter and the Democratic Necessity of Social Movements* (Oxford University Press).

About the Authors

Henry Farrell is the SNF Agora Professor of International Affairs at Johns Hopkins School of Advanced International Studies, and the 2019 recipient of the Friedrich Schiedel Prize for Politics and Technology. He is a member of the Council on Foreign Relations, and a Council Member of the European Council on Foreign Relations, as well as an affiliated scholar at Stanford University Law School's Center for the Internet and Society, and an international correspondent for Stato e Mercato.

Hahrie Han is the inaugural director of the SNF Agora Institute, the Stavros Niarchos Foundation Professor of Political Science, and the faculty director of the P3 Research Lab at Johns Hopkins University. She is an award-winning author of five books and numerous scholarly articles. She is an elected member of the American Academy of Arts and Sciences, was named a 2022 Social Innovation Thought Leader of the Year by the World Economic Forum's Schwab Foundation, and delivered the Tanner Lectures at Harvard University in 2024.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, policy advocacy, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org