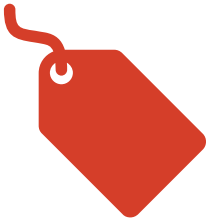

Ethics of
Social Media
Policy Project

Ethics Institute Working Paper:
https://ssrn.com/abstract=4805026

AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?

John P. Wihbey

Northeastern University
Ethics Institute

Northeastern University
College of Arts, Media and Design

AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?

John P. Wihbey, Northeastern University
College of Arts, Media & Design
Ethics Institute; Internet Democracy Initiative
j.wihbey@northeastern.edu

Conference on Democracy's Mega Challenges: How Climate Change, Migration, and Big Data Threaten the Future of Liberal Democratic Governance

Trinity College, Hartford, CT
April 19-20, 2024

Summary:

As advanced artificial intelligence (AI) technologies are developed and deployed, core zones of information and knowledge that support democratic life will be mediated more comprehensively by machines. Chatbots and AI agents may structure most internet, media, and public informational domains. What humans believe to be true and worthy of attention – what becomes public knowledge – may increasingly be influenced by the judgments of advanced AI systems. This pattern will present profound challenges to democracy. A pattern of what we might consider "epistemic risk" will threaten the possibility of AI ethical alignment with human values. AI technologies are trained on data from the human past, but democratic life often depends on the surfacing of human tacit knowledge and previously unrevealed preferences. Accordingly, as AI technologies structure the creation of public knowledge, the substance may be increasingly a recursive byproduct of AI itself – built on what we might call "epistemic anachronism." This paper argues that epistemic capture or lock-in and a corresponding loss of autonomy are pronounced risks, and it analyzes three example domains – journalism, content moderation, and polling – to explore these dynamics. The pathway forward for achieving any vision of ethical and responsible AI in the context of democracy means an insistence on epistemic modesty within AI models, as well as norms that emphasize the incompleteness of AI's judgments with respect to human knowledge and values.

Introduction

In a contemporary context, democracy can be seen as a system of information-processing that, ideally, aims to resolve problems through "collective cognition."[1] Any technology that fundamentally reorients the ways humans exchange and interact with information must be taken seriously as a potential threat, or at least a disruptive variable, in the context of democratic life. It is therefore no surprise that fears over the rise of generative artificial intelligence (AI) and its potential usurpation of human agency and democratic self-rule continue to draw attention from the public and policymakers around the world.

Many such issues in this problem space are categorized as potential "alignment" challenges – i.e., solving for the danger of AI operating without due reference to human values and preferences.[2] Some experts even worry about "existential risk" for humans.[3] However, scholars and technologists also continue to explore the ways in which AI, deployed correctly and thoughtfully, might actually improve deliberation, cooperation, collective intelligence, and decision-making across democracies, increasing the quality and flow of human preferences and inputs into public choices. This area of research has generally been characterized as the "mechanism design" challenge within the field of AI.[4]

Both the alignment and mechanism design research spaces will no doubt occupy a great deal of attention in the coming years as safety, reliability, and trustworthiness concerns continue to dominate public and expert debate.[5] Yet upstream from both problem areas is a matter that has largely escaped critical debate thus far: How are we to maintain citizens' organic education,

---

[1] Henry Farrell and Cosma Rohilla Shalizi, "Pursuing Cognitive Democracy" *From Voice to Influence: Understanding Citizenship in a Digital Age* (Chicago: University of Chicago Press, 2015) 211-231.

[2] Brian Christian, *The Alignment Problem: How can Machines Learn Human Values?* (London: Atlantic Books, 2021).; Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30, no. 3 (October 2020): 411-437. https://doi.org/10.1007/s11023-020-09539-2.; Russell, S. *Human Compatible: AI and the Problem of Control* (Bristol: Allen Lane, 2019).

[3] Andrew Critch and David Krueger, "AI Research Considerations for Human Existential Safety (ARCHES)," *arXiv preprint* (2020): arXiv:2006.04948.

[4] Raphael Koster, Jan Balaguer, Andrea Tacchetti, Ari Weinstein, Tina Zhu, Oliver Hauser, Duncan Williams et al., "Human-Centred Mechanism Design with Democratic AI," *Nature Human Behaviour* 6, no. 10 (July 2022): 1398-1407. https://doi.org/10.1038/s41562-022-01383-x.; Helene Landemore, "Fostering More Inclusive Democracy With AI," *Policy Commons*, Dec. 1, 2023, https://policycommons.net/artifacts/9768294/fostering-more-inclusive-democracy-with-ai-by-landemore/10657581/.

[5] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi, "Trustworthy Artificial Intelligence: A Review," *ACM Computing Surveys (CSUR)* 55, no. 2 (January 2022): 1-38. https://doi.org/10.1145/3491209.

value formation, and ongoing knowledge acquisition in the face of overwhelming AI saturation of informational domains? By "organic," I mean cognition that is not substantially shaped and influenced by machines. The alignment problem assumes that humans' values/preferences are (at least somewhat) organically derived and maintained, and it is the job of AI architects to come into conformity and express them, even as they may evolve. The mechanism design problem likewise assumes that, provided the technological platforms and models are structured properly, humans will be able to express their validly derived, organic views and have those reflected in democratic decisions.

For both alignment and mechanism design, it is logically necessary that AI does not itself substantially shape the very cognitive resources with which it aims to match; if it does, alignment, for example, would become recursive and incoherent, i.e., AI aligning with itself. AI systems that go too far in shaping human preferences, values, and what counts as knowledge – the cognitive resources "upstream," as it were – may ultimately undermine the very idea of alignment. The central danger becomes not misalignment, but rather corrupted or pseudo alignment.[6] For mechanism design, the risk is the same: If human preferences and reasoning are already a highly synthetic product of AI's shaping, then outputs from even a well-designed mechanism are hollow. The looming fundamental risk is epistemic capture or lock-in.

This paper addresses a central research question: What are the dangers to public knowledge and the concept of an informed citizenry in a democracy that is increasingly suffused with and mediated by advanced AI technologies? To approach this question comprehensively, it is necessary to analyze an interdisciplinary range of literature, theory, and research findings. My overall aim is to create a synthesized general overview of the risks and challenges of this coming era. The contribution intended here is meant to be in part synthetic, drawing on disparate disciplines, and normative in its recommendations.

After reviewing relevant historical context, applicable social science theory, and situating epistemic risk concerns in the general literature of the contemporary AI research field, the paper's analysis focuses on three example areas relating to human cognitive resources and public knowledge that may be impacted by generative AI: 1) journalism; 2) social media content moderation; and 3) polling. The analysis here proceeds to examine tradeoffs and risks in terms of citizens' ability to gain useful and accurate knowledge about the world and for that knowledge to be a public resource for collective sensemaking and decision-making. These example domains help us evaluate what we might want AI to help with in terms of facilitating collective cognition, and what zones we may want to preserve vigorously as human-centered areas of deliberation, ensuring a precious zone of human cognitive resources.

I argue here that much of the intellectual project of AI safety and ethical AI development – organized in many communities around the concepts of Fairness, Accountability, Transparency, and Ethics (FATE)[7] – is predicated on, indeed depends on, a zone of human knowledge and values – an area of vital cognitive resources – that remains in some sense outside of, or prior to, AI's influence. Striving for alignment and pursuing mechanism design, in other words, are necessary but not sufficient to achieve their intended program goals. This paper asserts that we must begin this era of mass engagement with AI technologies with a clear-eyed, and strongly normative, view of AI's limits – both what it cannot and should not do. The most plausible danger of the AI era may not be existential risk but rather a more subtle and incremental epistemic risk.

Any normative pathway forward for achieving any vision of ethical and responsible AI in the context of democracy requires instilling epistemic modesty within AI models and the creation of systems that emphasize the incompleteness of AI's judgments with respect to human public knowledge. Given the risks of recursion and the shaping of human cognitive resources "upstream" by advanced AI, democratic societies must evolve norms that place AI models in their proper context, as technologies that are always-already incomplete and lacking crucial data inputs that have yet to be revealed or created by humans and their communities. The very word "intelligence" in the context of AI must be used with extreme caution in the framing of applications that touch on democratic mechanisms and public knowledge.[8] There remains great hope that AI can in fact be a net-positive for facilitating greater democratic deliberation and including more voices in the conversation,[9] but the risk areas that could undermine such an optimistic vision have many subtle dimensions and in many ways are endemic.

---

[7] Bahar Memarian and Tenzin Doleck, "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and Higher Education: A Systematic Review," *Computers and Education: Artificial Intelligence* (June 2023): https://doi.org/10.1016/j.caeai.2023.100152.

[8] Precisely distinguishing forms and types of intelligence in this regard is a complicated matter. See: Christopher Newfield, "How to Make "AI" Intelligent; Or, The Question of Epistemic Equality," *Critical AI* 1, no. 1-2 (October 2023): https://doi.org/10.1215/2834703X-10734076.; Xiao-Li Meng, "Human Intelligence, Artificial Intelligence, and Homo Sapiens Intelligence?" *Harvard Data Science Review*, Issue 5.4 (October 2023): https://doi.org/10.1162/99608f92.11d9241f.

[9] Tantum Collins, "Democracy on Mars 3: New tools for popular sovereignty," Substack (Blog), Aug. 28, 2023, https://tantum.substack.com/p/democracy-on-mars-3-new-tools-for.; Helene Landemore, *Fostering More Inclusive Democracy With AI* (Washington, D.C.: International Monetary Fund, 2023): https://www.imf.org/en/Publications/fandd/issues/2023/12/POV-Fostering-more-inclusive-democracy-with-AI-Landemore.; Helene Landemore, "Can AI Bring Deliberative Democracy to the Masses?" In *HAI Weekley Seminar* (2022): 1-35; Bruce Schneier, Henry Farrell, and Nathan E. Sanders, "How Artificial Intelligence can Aid Democracy," *Slate,* April 21, 2023, https://slate.com/technology/2023/04/ai-public-option.html.

Power and Preferences

Even under ideal scenarios, some amount of "pure" or pre-AI human values, knowledge, and initial preferences are likely to be modified as AI technologies assume a greater role in the mediating and shaping of society. Within the general ethical framework of principles articulated across many expert and stakeholder institutions, such a modification of values, knowledge, and preferences implicates most directly the central principle of human autonomy.[10] Whether such loss of autonomy will lead to the outright takeover of knowledge and values over time remains to be seen, but it is imperative that AI designers and technologists attend to this problem of limits. They must be keenly aware that AI, in its systemic effects, may exercise softer, but in many ways deeper, forms of "power" by shaping initial preferences and structuring what people seek, desire, or believe to be valuable.

Social scientists have long noted the most entrenched categories of power involve changing, at the stage of initial preferences, what people think about and how they think about it.[11] This power ultimately becomes ideological in nature. As will be discussed, the diminution of human autonomy in this way may be profoundly damaging to democracy, which requires autonomous deliberation and choice-making. Even the desire to amplify human intelligence may clash with liberty and autonomous thought.

Of course, AI technologies are already shaping the media ecosystem broadly, leading to some degree of influence and shaping through algorithmic amplification, recommendations, and content production.[12] Scholars have noted a "recursive" dimension to the "data coil" that seems to spin in an endless loop powered by algorithmic curation and data mining in social media and other internet domains.[13] But as AI begins to become ubiquitous across new sectors and aspects of democratic human life, the question is how much such feedback loops may undermine the possibility of organic public knowledge and the degree of epistemic capture.

---

[10] Luciano Floridi and Josh Cowls, "A Unified Framework of Five Principles for AI in Society," *Machine Learning and the City: Applications in Architecture and Urban Design* (May 2022): 535-545. https://doi.org/10.1002/9781119815075.ch45.

[11] Steven Lukes, *Power: A Radical View* (London: Bloomsbury Publishing, 2021).; Joseph S. Nye Jr, *The Future of Power* (New York: PublicAffairs, 2011).

[12] Taina Bucher, *If... Then: Algorithmic Power and Politics* (Oxford: Oxford University Press, 2018).; Jenna Burrell and Marion Fourcade, "The Society of Algorithms," *Annual Review of Sociology* 47 (July 2021): 213-237. https://doi.org/10.1146/annurev-soc-090820-020800.

[13] David Beer, "The problem of researching a recursive society: Algorithms, data coils and the looping of the social," *Big Data & Society* 9, no. 2 (September 2022): https://doi.org/10.1177/20539517221104997.

In the context of democracy, the research community continues to surface new potential threats, including: disruption or warping of the communication pipeline between citizens and elected officials and regulators; threats to citizen trust due to synthetic content proliferation online; and threats to democratic representation, due to false waves of signals to political systems and officials.[14]

The coming AI era may or may not be necessarily fatally flawed, doomed to usher in the development of AI supremacy over humans and "algocracy" or, at the edges of speculation, mark the extinction of the human species.[15] Experts continue to predict that AI will substantially reshape nearly all elements of human life.[16] If there is to be any chance of success for integrating these technologies into society while preserving genuine democracy, society will need to think carefully about how to preserve certain human cognitive resources and value formation, in order to maintain a wellspring of organic deliberation and understanding among humans, ungoverned or optimized by AI systems.

Informed Citizenry and Model Deficits

Scholars from across the social sciences and humanities have long maintained that an informed citizenry is necessary to achieve any potential vision of a well-functioning democratic society.[17] Without knowledge, citizens cannot participate fully in their polity; the demos cannot make wise choices in terms of choosing leaders and allocating resources. Without informed choice-making capacity, a would-be democratic society is subject to manipulation, corruption, and domination. Democracy may become a sham without the informed citizen.

Yet there has long been a skeptical tradition that maintains it is unrealistic that citizens will be "omnicompetent," and tests of political knowledge and the relative engagement levels of citizens – and studies of their ability to make choices aligned with their interests – in

[14] Sarah Kreps and Doug Kriner, "How AI Threatens Democracy," *Journal of Democracy* 34, no. 4 (October 2023): 122-131.
[15] John Danaher, "The Threat of Algocracy: Reality, Resistance and Accommodation," *Philosophy & Technology* 29, no. 3 (January 2016): 245-268. https://doi.org/10.1007/s13347-015-0211-1.
[16] Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake our World* (New York: Basic Books, 2015).; Mustafa Suleyman, *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma* (New York: Crown, 2023).
[17] Danielle Allen and Rob Reich, eds. *Education, Justice, and Democracy* (Chicago: University of Chicago Press, 2019).; Josiah Ober, "From Epistemic Diversity to Common Knowledge: Rational Rituals and Cooperation in Democratic Athens," *Episteme* 3, no. 3 (2006): 214-233. https://doi.org/10.3366/epi.2006.3.3.214.; Thomas E. Patterson, *Informing the News: The Need for Knowledge-Based Journalism* (New York: Vintage, 2013).

democracies suggest that reality often falls well short of ideal theory.[18] As the late Russell Hardin once wrote, the evidence stands heavily on the skeptics' side: "The data are merciless and brutal. The typical voter is politically ignorant and often misguided."[19] Further, a large body of research relating to human cognition suggests that bias and fallibility in judgment are permanent challenges to a citizen's rationality in belief and decision-making.[20]

Such facts about human ignorance and irrationality fuel the constant temptation to supplant democracy with greater rule by elites, intrusive central planning, oligarchy, and various other forms of anti-democratic rule. It is not hard to imagine that the AI era will fuel more such temptations to impose top-down influence and rationalizing models on human public choices. In response to criticisms of democracy as a system, political theorists have argued that the point of democracy is not perfect education or rationality but rather communication and deliberation.[21] We might judge a democratic system's quality, however messy, by its ability to bring people into reasoning and deliberating opportunities, giving the system legitimacy and authority.[22] In the internet era, we might also judge systems further by their capacity to ensure inclusion and cross-cutting conversation.[23]

Technological shifts have added new wrinkles in the long-running debate over the informed citizen. The era of relatively centralized mass broadcast in the twentieth century saw media and political communication scholars become increasingly alarmed at the prospect of the democratic public being seduced by entertainment and soundbite culture, superficial distractions from the hard business of learning, informing, and performing democratic self-

---

[18] Christopher Achen and Larry Bartels, *Democracy for Realists: Why Elections do not Produce Responsive Government* (Princeton, NJ: Princeton University Press, 2017).; William A. Galston, "Political Knowledge, Political Engagement, and Civic Education," *Annual Review of Political Science* 4, no. 1 (June 2001): 217-234. https://doi.org/10.1146/annurev.polisci.4.1.217; "What the Public Knows—In Pictures, Words, Maps and Graphs," Pew Research Center, April 28, 2015, https://www.pewresearch.org/politics/2015/04/28/what-the-public-knows-in-pictures-words-maps-and-graphs/.

[19] Russell Hardin, "Deliberative Democracy," in Thomas Christiano and John Christman, *Debates in Political Philosophy,* (Wiley Online Library, 2009): p. 229.

[20] Daniel Kahneman, *Thinking, Fast and Slow* (New York: Macmillan, 2011).; Richard H. Thaler and Cass R. Sunstein, *Nudge: The Final Edition(* New Haven, CT: Yale University Press, 2021).

[21] John Dewey, *The Public and its Problems* (Athens, OH: Swallow Press, 1954): First published 1927 by Henry Holt and Company (New York).

[22] Joshua Cohen, "Procedure and Substance in Deliberative Democracy," in James Bohman, William Rehg, eds. *Deliberative Democracy: Essays on Reason and Politics (*Cambridge, MA: The MIT Press, 1997).

[23] Cass Sunstein, *# Republic: Divided Democracy in the Age of Social Media* (Princeton: Princeton University Press, 2018).

rule.[24] Critical theorists also worried that the vehicles of radio and television would simply reinforce power structures, and what might seem like an informed citizenry might ultimately prove to be a bovine mass public indoctrinated by corporations, interest groups, and government propaganda.[25] In this era, the structure of media industries and the quality of media were central areas of focus and debate.

Since the advent of the interactive and social web (which we might date to, say, the founding of Facebook in 2004), there has been growing concern that citizens are being misinformed — and collective cognition therefore corrupted — by a decentralized deluge of half-truths, lies, false narratives, misinformation, disinformation, networked propaganda and a ream of other informational forms that are distinctive of this era. The chief fear is that the "informing" part of democratic life is warped by a combination of malign actors and algorithms that replace high-quality information with corrupted or distracting information.[26] Proposed solutions have varied, running from: social media companies tuning algorithms toward quality information; media literacy in civil society; increased efforts at content and source labeling; downranking or decreasing the visibility of low-quality information; removal of corrupted information; and inoculation and pre-bunking strategies by social platforms. The focus in this era has been on better protecting the integrity of the citizen's mind against intrusive forces and on regulating the decentralized platforms that filter and amplify content damaging to truth and democracy, as well as public health and other important domains.

Given these prior waves of media history – of information "mediation" history – the emerging era defined by generative AI presents a novel context in which profound challenges arise for the concept of the informed citizen.

Reinforcement and Rewards

[24] Neil Postman, *Amusing Ourselves to Death: Public Discourse in the Age of Show Business* (London: Penguin, 2005).
[25] Edward S. Herman and Noam Chomsky, *Manufacturing Consent: The Political Economy of the Mass Media* (New York: Pantheon, 1988).
[26] Eytan Bakshy, Solomon Messing, and Lada A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science* 348, no. 6239 (May 2015): 1130-1132. 10.1126/science.aaa1160.; Ariadna Matamoros-Fernández, Joanne E. Gray, Louisa Bartolo, Jean Burgess, and Nicolas Suzor, "What's "Up Next"? Investigating Algorithmic Recommendations on YouTube Across Issues and Over Time," *Media and Communication* 9, no. 4 (November 2021): 234-249. https://doi.org/10.17645/mac.v9i4.4184.; Eli Pariser, *The Filter Bubble: How the New Personalized Web is Changing What we Read and how we Think* (London: Penguin 2011).

The AI research community has continued to work on a variety of areas that loosely revolve around what has traditionally been called the "principal-agent" problem[27] – in this case, the ability of an AI "agent" to follow/match with the preferences of a human "principal." We can evaluate and debate AI systems at the more abstract level of general alignment and model-level objective functions (goals toward which the system aims), but it is useful to begin at the more fundamental and technical level of principal-agent interaction, which then ladders up to more general systemic issues.

Generative AI models are continually trained on vast amounts of data from the human-created web. A central problem that model training faces is "overfitting," where a model trained on one dataset is unable to apply its learning to a new dataset.[28] The model initially learns in such a way that it cannot generalize to a larger domain. Overfitting often happens when training data are too narrow or lack relevance.

AI models often undergo fine-tuning through a process called "reinforcement learning."[29] Further refinement often involves the technique of "reinforcement learning from human feedback" (RLHF), which is the practice/area of AI research that concerns the extension of large language models (LLMs) to areas of uncertainty for the models. Such processes are used to produce OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude, among others. RLHF techniques "allow LLMs to go beyond modeling the distribution of their training data, and adapt the distribution of text so that model outputs are rated more highly by human evaluators."[30] RLHF involves LLMs making probabilistic guesses about novel areas of knowledge (e.g., expressions, terms, concepts, arguments, inferences) and paid human raters providing judgments about the success of those guesses. These evaluations then help train the model to become better at extending its base of knowledge to new areas. This is sometimes called

[27] Dylan Jasper Hadfield-Menell, *The Principal–Agent Alignment Problem in Artificial Intelligence* (Berkeley, CA: University of California, Berkeley, 2021).; Michael Jensen, and William H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of Financial Economics* 3, 4 (October 1976): 305–360. https://doi.org/10.1016/0304-405X(76)90026-X.; Kathleen M. Eisenhardt, "Agency Theory: An Assessment and Review," *The Academy of Management Review* 14, 1 (January 1989): 57–74. https://doi.org/10.1016/0304-405X(76)90026-X.

[28] Xue Ying, "An Overview of Overfitting and its Solutions," In *Journal of physics: Conference series*, vol. 1168 (February 2019): p. 022022. https://dx.doi.org/10.1088/1742-6596/1168/2/022022.; "What is Overfitting?" Amazon Web Services, n.d., https://aws.amazon.com/what-is/overfitting/ .

[29] Richard S. Sutton and Andrew G Barto, *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press, 2018).

[30] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, et al. "Open Problems and Fundamental Limitations of Reinforcement Learning From Human Feedback," *arXiv preprint* (September 2023): *arXiv:2307.15217,* 2.

creating an accurate "reward function" computationally, translating from what the human rater specifies to what the algorithm ought to do in future cases.

However, as scientists working at the leading edge of reinforcement learning research have noted, "reward functions" are often unable to express the complexity of human thought and values; moreover, the doubtful prospect of reward functions representing a diverse society, with diverging values that often evolve over time and are sometimes contradictory, is a fundamental challenge to aspirations of producing universally valid models.[31]

Still, it is necessary to get the training of these systems right, to the extent possible. AI systems that are not properly trained will have significant downstream and tangible consequences for democratic societies, some of which are not long-term and speculative but rather are unfolding right now.[32] Such problems may include bias and discrimination in automated decision-making systems, affecting economic-, health-, elections-, and justice-related domains, as well as the potential spread of mis- and dis-information across digital platforms by malign actors leveraging AI.

Persistent Model Deficits

There is a fundamental irony as we enter the AI era, which is that public knowledge may be both enhanced and undermined simultaneously. Under the most charitable futuristic scenarios, generative AI models might serve as knowledge-bearing "agents" and co-pilots, greatly empowering humans' ability to access and leverage knowledge, or systematic information. The generative AI era should, in theory, create a more informed citizen than was possible in any prior era in history. The answer to most any factual question, the proper approach to an illness, the ideal travel itinerary in Peru, the effect of a budget hike on taxes – on such issues, generative AI might instantly guide humans better than they could guide themselves, assuming such learning technologies continue to improve rapidly over the coming decades in terms of accuracy and contextual understanding. Further, the rapid creation and application of collective intelligence – combinations of human knowledge to address problems and challenges — seems much more possible, given the ability of AI technologies to solve matching problems and

[31] Casper, et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback," 8-9.

[32] Andreas Jungherr, "Artificial Intelligence and Democracy: A Conceptual Framework," *Social Media + Society* 9, no. 3 (July 2023): https://doi.org/10.1177/20563051231186353; Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel, "Open Problems in Cooperative AI," *arXiv preprint* (December 2020)*:* arXiv:2012.08630.

synthesize large, nuanced bodies of information.[33] That is the case for public knowledge enhancement – the optimistic case, as it were.

As for the idea of knowledge being undermined: What, exactly, might be the outlines of the coming problems? In some ways, the very structure of artificial intelligence technologies, namely the building of predictive models based on past data (backward-looking training data), is fundamentally in conflict with basic aspects of democratic life, which is inherently forward-looking. Democracy is fundamentally emergent; AI models are epistemically anachronistic. There are latent aspects of human democracy that are seldom considered in the true extent of their importance, in part because society has never had to contemplate the replacement or short-circuiting of them. AI models will always carry a kind of epistemic risk, as they are structurally missing vast amounts of data about what humans actually consider to be true, important, useful, and interesting. AI technologies may, by their very nature, attempt to structure and constrain the creation of signals from citizens. To put this argument in the language of technology, citizen participation in public life requires the constant, organic creation and activation of "new data" and new signals or variables; this data is necessarily missing data in any AI model, because those datapoints do not yet exist as public information susceptible to model classification and aggregation.

Of course, reinforcement learning techniques might try to help overcome this structural deficit of epistemic anachronism, by providing guidance on future extensions of the model. But the implications over time for democratic systems are much deeper than has perhaps been considered by AI researchers.

The problems that AI presents for democracy in this regard are multifaceted, and we might consider them across five dimensions: First, humans rely on an enormous amount of tacit knowledge.[34] Much of what is known or believed by citizens is not yet expressed — indeed, we know more than we can say — and AI models are bereft of vast stores of crucial information, as models are trained on data from the existing web. Second, citizens often falsify preferences because of social desirability bias and only reveal true preferences as a result of visible cascading social norms[35]. Third, humans come to almost all of their knowledge through other humans and social contact; this phenomenon has in recent decades generated an entire branch

[33] Dafoe et. al."Open Problems in Cooperative AI."
[34] Michael Polanyi, *The Tacit Dimension* (Chicago: University of Chicago, 2009).; Michael Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy (*Milton Park, Abingdon, Oxon*:* Routledge, 2012).
[35] Cass R. Sunstein, *How Change Happens* (Cambridge, MA: MIT Press, 2019).

of philosophy called "social epistemology."[36] Fourth, citizens' sense of agency and liberty – their sense of whether an action, rule, or system is fair, just, freedom-preserving or desirable – is often bound up with spontaneous interactions between real-world events and important, non-rational elements of human cognitive experiences such as emotions, intuitions, and acts of imagination. Fifth, as many political theorists contend, it is through the exercise of citizens' use of reasoning faculties together in human groups that citizens authentically participate in an inclusive model of "deliberative democracy."

One overarching issue affecting all domains that AI touches is that feedback loops will be an inherent feature in the process of deliberation, insofar as epistemically anachronistic AI agents trained on past data will influence future deliberation in ways that prevent new knowledge from developing organically by humans. The danger from such feedback loops is epistemic capture or lock-in — what is considered true, interesting, and useful by humans is continually mediated by AI, which then reinforces past ideas and preferences. A consequence of capture, one profoundly affecting democratic life, is that humans may not be able to access emerging, organic ground-truth signals from fellow citizens about their dispositions on issues, potentially silencing important intuitions, emotions, tacit knowledge, gut feelings, and experiences.

Risk Domains

Let us examine three example areas relating to public knowledge, and thus the concept of informed citizenry, that may be impacted by generative AI: 1) journalism; 2) social media content moderation; 3) polling. In each of these cases, there has been an existing institutional mediating force between the citizen and knowledge/information. AI presents a new form of mediation that, in many cases, may be transformative. This impending change should prompt a deep examination of tradeoffs and risks in terms of citizens' ability to gain useful and accurate knowledge about the world and for that knowledge to be a public resource for collective sensemaking and decision-making.

In the case of 1) journalism, there are already a number of experiments underway to try to generate automated coverage of events, including civic and government meetings, leading to the possibility of entire municipalities or even larger regions being covered by AI news agents. What remains unclear is how such automated processes will fulfill recent aspirations for "networked journalism" and facilitation of co-created sense-making in communities. In the second case, that of 2) content moderation in social media, there is the distinct prospect of AI-powered moderator agents, developed and deployed by social media companies themselves,

[36] Alvin I. Goldman, *Knowledge in a Social World* (Oxford: Oxford University Press, 1999).

intervening in public conversations to enforce rules, provide assistance, or fact-check/provide context – "modbots" or "chatmods," as it were. In the case of 3) polling, there are likely to be increasing attempts to create synthetically produced dynamic models of public opinion by, in effect, setting up panels of interactive agents, trained on datasets that approximate human preferences, who stand in for human poll respondents. (Some are referring to this as "silicon sampling.")[37] Current research on AI models and their value for predictive polling and opinion estimates suggest that epistemic anachronism is a particularly acute problem. Emerging issues and predictions of public reception prove inherently difficult for AI models, pointing to the structural problem of epistemic risk across all informational zones that affect democratic societies and their deliberative capacities.

Risk Area 1: Journalism

Journalism has long aspired to help provide public, orienting pictures of shared reality that support modern societies with collective decision-making.[38] Even while critics have often pointed out the inadequacies of journalism – from bias and inaccuracy to sensationalism and innumeracy – few doubt its importance as a central sense-making and coordinating resource for deliberation and debate. Some scholars have called for journalism to become more systematic and knowledge-based, particularly given the vast new possibilities for mis- and dis-information and manipulation afforded by internet technologies.[39] Meanwhile, the business model for newspaper and digital news journalism in particular has come under severe stress in recent years because of the flight of advertising revenue to large internet platforms, creating what many are calling a crisis for democracy in countries such as the United States.[40]

Those studying the intersection of AI and journalism have noted the wide variety of experiments underway to incorporate AI technologies into nearly every aspect of newsroom

[37] Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim, "Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information," *arXiv preprint* (February 2024)*:* arXiv:2402.18144.

[38] Bill Kovach and Tom Rosenstiel, *The Elements of Journalism, Revised and Updated 4th edition: What Newspeople Should Know and the Public Should Expect* (New York: Crown, 2021).

[39] Jean Folkerts, John Maxwell Hamilton, and Nicholas Lemann, *Educating Journalists: A New Plea for the University Tradition* (New York: Columbia University Graduate School of Journalism, 2013): https://www.journalism.columbia.edu/system/documents/785/original/75881_JSchool_Educating_Journalists-PPG_V2-16.pdf.; Patterson, *Informing the News*.

[40] Martha Minow, *Saving the News: Why the Constitution Calls for Government Action to Preserve Freedom of Speech* (New York: Oxford University Press, 2021).; Margaret Sullivan, *Ghosting the News: Local Journalism and the Crisis of American Democracy* (New York: Columbia Global Reports, 2020).

workflow, including "information discovery, verification, and content categorization, to enable large-scale analysis of social media and news coverage, to monitor public interest in specific topics, or to facilitate various kinds of (investigative) reporting."[41] New ethical frameworks are urgently needed as AI becomes deeply intertwined with the journalistic process.[42]

Over generations, computer software has brought automation to nearly every function within journalism except the independent creation of accurate, descriptive human language and its higher-level derivatives – from story ideas and frameworks for coverage to nonfiction narrative material itself. LLMs have begun to conquer this last frontier of language and story generation within journalism. While models still have significant drawbacks relative to expert journalists, they have proved capable of producing passable stories in data-driven and structured domains such as finance, sports, and elections. They are massively creative in writing headlines, formulating new questions about any topic, and generating framings for approaching subjects. The acceleration of AI technologies in all of these regards suggests that rapid progress can be expected, even as humans must remain in the loop for highly complex stories and for final fact-checks/editing to maintain quality and avoid costly errors that could hurt news brand reputation.

Given this context, let us extend these trendlines and consider a future scenario: A region of the United States has become a bonafide "news desert," and lacking any traditional institutions to perform journalism, a group of entrepreneurs have launched an AI news outlet that will "cover" the state capitol as well as many dozens of surrounding communities.[43] The news outlet has minimal staff – just enough, perhaps, to fact-check stories involving high-profile persons such as the governor or mayor. Otherwise, an AI agent, trained on data from all prior stories ever written or produced about the region, scrapes the web for news of government meetings and reports, law enforcement activities, sports, weather, and emergency services-involved events. It even deploys tailored emails asking for human comment on news events, performing basic interview functions and furnishing quotations from humans. The news produced for the region

[41] Felix M. Simon, and Luisa Fernanda Isaza-Ibarra, *AI in the News: Reshaping the Information Ecosystem?* (Oxford: Oxford Internet Institute, 2023): p. 8.
[42] Nicholas Diakopoulos, Hannes Cools, Charlotte Li, Natali Helberger, Ernest Kung, and Aimee Rinehart, "Generative AI in journalism: The evolution of newswork and ethics in a generative information ecosystem," Associated Press, 2024. https://www.aim4dem.nl/out-now-generative-ai-in-journalism-the-evolution-of-newswork-and-ethics-in-a-generative-information-ecosystem/.; Sachita Nishal, "Blueprints for Evaluating AI in Journalism: Generative AI in the Newsroom," *Medium*, March 28, 2024, https://generative-ai-newsroom.com/blueprints-for-evaluating-ai-in-journalism-e702c9e8c4f3.
[43] Penelope Muse Abernathy, *The Rise of a New Media Baron and the Emerging Threat of News Deserts* (University of North Carolina School of Media and Journalism: Center for Innovation & Sustainability in Local Media, 2016). http:// newspaperownership.com/.

turns out to be relatively reliable, even if a bit dull. The AI agent, programmed to avoid bias and emphasize fairness, consistently looks for outlier stories and attempts to quote voices from various groups who are not well represented.

In terms of fostering an informed citizenry and democracy, what do we make of such a scenario? Might it fulfill some minimum standard for meeting what has been called the "information needs of communities"?[44] Analytically, we could evaluate such an experiment on the quality of the output along a number of traditional lines in journalism and communication studies: its factual accuracy; the alignment of its framing of stories and its agenda-setting with respect to the needs of citizens; the ability of such AI news to report on powerful institutions and hold them accountable.[45] We could also factor in cost-benefit analysis. The region had no real news prior, so the creation of some amount of news, at a low and sustainable cost, might be a net-positive, even if diminished in overall creativity. Perhaps AI offers a minimum viable product for news.

The point is that this all could be done, theoretically, although we should acknowledge that media ethicists have grave concerns about the automated generation of entire pieces given the current state of technology.[46] Here we must return to the ideas of alignment and mechanism design in AI. Once the AI news outlet model has begun churning out stories, it has begun a feedback loop. It has begun doing agenda-setting, framing of narratives, and selection of what's important in a community. The very communities it covers might consume and engage with the material. The behavior of the actors in the community then adjust their preferences and decision-making accordingly. It is quite possible that the organic signals of humans become ever-more-difficult to detect, uninfluenced by the media environment that the AI has created. The model's process of tail-eating has begun. The model, of course, could look for novelty – previously unrevealed preferences, emotions, intuitions, and yet-unexpressed ideas from citizens.

It is at this point that the media critics might raise an obvious point: Human journalists have long been biased, or at least highly susceptible to using predictable framings, and there are generations of scholarship that bear out this point. But that is in some ways beside the point, as

[44] Steven Waldman and the Working Group on Information Needs of Communities Federal Communications Commission, *Information Needs of Communities: The changing media landscape in a broadband age* (Durham, NC: Carolina Academic Press, 2011).
[45] There are a variety of technical approaches to such evaluation of news quality. See: Philip Napoli, M., Sarah Stonbely, Kathleen McCollough, and Bryce Renninger, "Local Journalism and the Information Needs of Local Communities: Toward a Scalable Assessment Approach," *Journalism Practice* 11, no. 4 (August 2019): 373-395. https://doi.org/10.1080/17512786.2016.1146625.
[46] Diakopoulos, et al. "Generative AI in Journalism."

the central question remains the ethical alignment of machine technologies in democratic human societies.

The obvious solution is to keep humans in the loop, allowing the AI news agent to continue to get human signals and provide access to values and preferences that are not that of the machine. Yet that will require a very deliberate effort to wall off a zone of human cognitive resources that are not themselves the byproducts of AI shaping. It will also require some strong normative preference for ideas, values, preferences, and interactions that directly derive from humans, and are not mediated by AI.

A final critique is that theorists of journalism have increasingly noted the importance of social connection through the newsmaking process.[47] The idea is that news and newsmaking are a kind of process and platform that allows citizens to deliberate together, with the journalist facilitating, ideally speaking, "collaborative intelligence."[48] Social media have made this function of journalism quite explicit. It is in this way that AI is additionally problematic, as it plays a role in social media space, through algorithms, that mediate human relationships. Again, alignment becomes potentially incoherent when the signals for mediation are derived from feedback loops — based on the reward functions in its model, AI has mediated the news material, the agenda, and the social ties.

The AI journalism scenario points to three principles that can be carried through to examine AI's effects across other adjacent, example domains: 1) All AI models are inherently in danger of being contradictory from an alignment perspective with respect to public knowledge; 2) There must be a zone of human cognitive resources separate from heavy AI intervention; 3) Not only do humans individually need space to formulate preferences, knowledge, and values, but they must have space to connect with other humans, in ways that are not entirely steered by the reward functions in the AI models.

The norms of the traditional, human news business dictate crisp, definitive accounts of human actions, events, and views – blaring headlines and simple, condensed summaries that attempt to declare empirical truth and certain judgment. Should AI attempt to mimic such norms? There

[47] John P. Wihbey, *The Social Fact: News and Knowledge in a Networked World* (Cambridge, MA: MIT Press, 2019).
[48] Tom Rosenstiel, "News as Collaborative Intelligence: Correcting the Myths About News in the Digital Age," Center for Effective Public Management at Brookings, June 30, 2015, https://www.brookings.edu/research/news-as-collaborative-intelligence -correcting-the-myths-about-news-in-the-digital-age/.

are already attempts to create synthetic, "deep fake" news anchors;[49] as mentioned, thousands of news articles are being produced by AI, in the exact forms, grammar, and idiomatic modes of expression of human journalists, even as there remain serious questions about models' ability to do this well.[50]

If AI assumes large areas of the societal function of news media-making, it will be well worth attending to issues of epistemic humility in AI models, and how AI might express such modesty in the way it delivers news. One mitigating idea, taken from what has been called "human-centered AI," might be to more clearly delineate the behavior and products of humans and machines, and shifting away from emulating humans and their work.[51] Human-centered AI theories insist that AI bots and agents should very clearly distinguish themselves in their interactions with humans. An area of further research within news production might revolve around related questions – how AI-centered news production can look, sound, and feel quite different from that of human-generated news, foregrounding its origins and conveying humility and incompleteness.

Risk Area 2: Content Moderation

Every day on the social web, millions of user-generated posts kick off a long, winding debate about some matter of public concern. A citizen posts her thoughts on immigration policy, for example, on a social platform, such as Instagram, YouTube, X, or Telegram. It is a bit edgy in its language. The post spurs intense reactions; emotions run high. Soon, the inevitable happens, and an online participant steps over the line, violating community standards of whatever social platform is being used. The platform's moderators step in. But what if increasingly, that moderator is a company-sponsored AI agent?

Social media platforms typically moderate content by downranking it, taking it down, or providing warning labels – a framework referred to commonly as "reduce, remove, inform."[52]

---

[49] Matyas Bohacek and Hany Farid, "The making of an AI news anchor—and its implications," *Proceedings of the National Academy of Sciences* 121, no. 1 (December 2023): https://doi.org/10.1073/pnas.2315678121.

[50] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. "Bias of AI-generated content: an examination of news produced by large language models." *Scientific Reports* 14, no. 1 (March 2024): 1-20, https://doi.org/10.1038/s41598-024-55686-2.

[51] Ben Shneiderman, *Human-Centered AI* (New York: Oxford University Press, 2022).

[52] Tarleton Gillespie, "Do Not Recommend? Reduction as a Form of Content Moderation," *Social Media + Society* 8, no. 3 (August 2022): https://doi.org/10.1177/20563051221117552.; Tessa Lyons, "The Three-Part Recipe for Cleaning up Your News Feed," Meta, May 22, 2018, https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/.

These actions are often powered by machine learning and AI. Now, a new potential approach looms on the horizon: "chatmods," or "modbots" (chatbot moderators) – company-deployed interactive AI agents that can help referee or police rules in direct dialogue with users.[53] These generative AI tools may potentially be deployed for, among other things, information assistance, mediation, warnings, network disruption, addressing disinformation efforts through counterspeech, or even prebunking or inoculation efforts on platforms.[54]

Of course, a huge amount of AI and machine learning structure existing algorithms to reduce, remove, and inform; automated decision-making powered by learning technologies suffuses social media, and has so for more than a decade. But large social media companies have so far not rolled out dynamic, and riskier, moderation approaches that might allow users to engage in conversation, explain themselves, plead their case, or allow others in to see moderation actively taking place via an embodied "referee" bot on the platform. Social media users, of course, are accustomed to seeing bots play a role in the information environment, but these are not company-sponsored, with formal authority and power to take actions as they engage users.

There has been considerable discussion about the intentions of companies deploying LLMs, such as OpenAI, to help further with content moderation.[55] Observers in the trust and safety field have grave concerns about the readiness and capability of existing models to do high-quality, effective, and ethical work.[56] Companies have experimented with creating publicly accessible AI personalities, such as Meta/Facebook's Jane Austen bot, while the social media company Snap offers a "My AI" chatbot to users.[57] Discord and Reddit have seen experiments with automated moderator technologies within those platforms' more decentralized communities.

---

[53] John Wihbey and Garrett Morrow, "Social Media's New Referees?: Public Attitudes Toward AI Content Moderation Bots Across Three Countries," Ethics Institute, Northeastern University, 2024.
[54] Sayash Kapoor and Arvind Narayanan, "How to Prepare for the Deluge of Generative AI on Social Media," Knight First Amendment Center, Columbia University, June 16, 2023, https://knightcolumbia.org/content/how-to-prepare-for-the-delugeof-generative-ai-on-social-media
[55] AV Lillian Weng, Vik Goel, and Andrea Vallone, "Using gpt-4 for content moderation," OpenAI, August 15, 2023, https://openai.com/blog/using-gpt-4-for-content-moderation.
[56] Alyssa Boicel, "Using LLMs to Moderate Content: Are They Ready for Commercial Use?" *Tech Policy Press*, April 3, 2024, https://www.techpolicy.press/using-llms-to-moderate-content-are-they-ready-for-commercial-use/.
[57] "What is My AI on Snapchat and how do I use it?" Snapchat.com, Accessed April 4, 2024, https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it.

Any top-down chatmod strategy currently would be risky, given that the models carry racial, gender, and cultural bias; LLMs are "non-deterministic," insofar as companies cannot entirely predict their reaction to human engagement.[58] To date, such technologies also have not demonstrated the ability to fact check claims at a reasonably high level.[59]

A 2023 study of public opinion, led by this author, across the United States, the United Kingdom, and Canada found consistent worry about the risks of companies using chatbots to perform content moderation in social media. A majority of survey respondents in all three countries expressed concern that AI chatbots would misunderstand the context of users' posts, would make flawed decisions, could ruin the experience of interacting with other humans, might introduce false information, and might make social media more divisive.[60]

Setting aside the ability of chatbots to perform the job reasonably well – the functionality problems that remain formidable – there is a larger question about the potential loss of social and knowledge value for democracy as machines begin assuming the tasks of refereeing public conversations over issues in the digital public square. What would be lost? The answer, potentially, is public knowledge – about what people think, how they argue, the norms they battle for, and the preferences that they reveal as they debate and deliberate. For sure, well-functioning AI chatmods might provide speed and scale to an often-overwhelming task of keeping hate speech, disinformation, and the like from spinning out of control. But what could be lost is a great deal of socially and democratically useful information and knowledge, as users argue back-and-forth. There is value in witnessing corrections – in experiencing reasoning processes and debate – by other humans.[61] Experts on hate speech see great value, as well, in digital bystanders sticking up for others and participating in the process of pushing back against antisocial voices, as such situations help to reinforce prosocial norms.[62]

Given the fluidity of human language and meaning-making – and, as discussed, the emergent nature of much human thought and ideation – there is also the question of how AI chatmods in

[58] Emilio Ferrara, "Should Chatgpt be Biased? Challenges and Risks of Bias in Large Language Models," *arXiv preprint* (November 2023): arXiv:2304.03738.
[59] Matthew R. DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer, "Fact-checking information genereated by a large language model can decrease news discernment," *arXiv preprint* (December 2023): https://doi.org/10.48550/arXiv.2308.10800.
[60] Wihbey and Morrow, "Social Media's New Referees?"
[61] Leticia Bode and Emily K. Vraga, "Correction Experiences on Social Media During COVID-19," *Social Media+ Society* 7, no. 2 (April 2021): https://doi.org/10.1177/20563051211008829.
[62] Susan Benesch and Cathy Buerger, "Can AI Rescue Democracy? Nope, It's Not Funny Enough," *TechPolicy Press*, March 11, 2024, https://www.techpolicy.press/can-ai-rescue-democracy-nope-its-not-funny-enough/.

this context could reasonably be expected to make ethical decisions that do not cause potential harm. Alignment in such situations is inherently difficult given the emergent nature and context of much online discourse, where breaking events and novel ideas often first surface.

Social media companies assert their corporate and fiduciary rights to keep their platforms safe and trustworthy. But as a chatmod engages in public or private with a user, serious questions arise about the ethics of such human-machine interactions.

Many AI ethics codes have extended the standard principles underpinning bioethics: beneficence; non-maleficence; autonomy; and justice. Some have also added explainability, or the obligation to be able to articulate reasons for decisions or generally how things work.[63] AI technologies are already significantly impacting users in all kinds of more invisible or hardly visible ways, from the decisions by what are termed "classifiers" that are not taken (false negatives) to over-enforcement (false positives.) It is not clear that AI as deployed currently is at all conforming to ethical principles. The harms created by current content moderation regimes are well-documented, raising profound ethical questions.[64]

Yet the use of machine bot personas to intervene in social space in a more visible and interactive way presents new ethical considerations, raising questions about the treatment of humans by machines. Even if a human user is posting messages that are borderline abusive, what right, and under what guidelines, does a chatmod operate under as it pushes back on a human's expression? Does the action embody the beneficence, non-maleficence, autonomy, justice, and explainability principles? All of this will depend very much on the behavior and the responses of the chatmod, as well as the policy framework used by the underlying platform. The decision to use a chatmod generates further ethical obligations on behalf of the content moderator system. This is an area greatly in need of more research from a human-computer interaction (HCI) and applied ethics perspective. But as a baseline, it would be reasonable to expect that social platforms: a) emphasize the clear identification of the agent as non-human in form (consistent with previously mentioned human-centered AI principles); and b) instill the concept of epistemic humility by the AI agent.

The use of persona-like agents in social space may just be the beginning of the widespread use of AI-powered robots in human life, a vision that has long entranced science fiction creators. What must be remembered is that all social interactions within democratic life have the

[63] Floridi and Cowls, "A Unified Framework of Five Principles for AI in Society."
[64] Sahana Udupa, Antonis Maronikolakis, and Axel Wisiorek, "Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence." *Big Data & Society* 10, no. 1 (May 2023): https://doi.org/10.1177/20539517231172424.

potential to create useful human-to-human deliberation and learning. In an ideal sense, social media companies might structure their platforms and moderation approaches to maximize the building of a healthy, high-quality information environment and to help users better navigate this environment.[65] Chatmods may advance such a mission in some respects. But human-to-human interactions also contain intrinsic value, as social ties can help powerfully transmit information, norms, and values. As AIs erase or manipulate such opportunities for humans to debate, disagree, connect, and interact, social platforms may cut short aspects of deliberative democracy that are vital.

Risk Area 3: Polling

News and social media are perhaps the two most expansive and important areas of potential democratic epistemic risk, insofar as these zones of public knowledge, information, and debate dominate the contemporary public squares of democracies around the world. However, the area of public opinion measurement and polling offers perhaps the most crisply defined of these three case studies in terms of highlighting epistemic risk. Issues of bias in AI models can become self-reinforcing as flawed pictures of public opinion influence public knowledge and human belief. Such flawed pictures may be generated by models that contain biases and by the problem of epistemic anachronism.[66] Such distortions in the information environment may affect democratic representation, accountability and trust.[67]

Numerous teams of researchers have experimented with using large language models as potential test beds for measuring public opinion. As might be expected, issues of bias and epistemic anachronism make predictions exceedingly difficult in cases that are not established in the model's training data. Polling experts warn that "big data" and AI-driven models can create more problems than they may solve, and they may threaten democracy.[68]

---

[65] John Wihbey, Matthew Kopec, and Ronald Sandler, "Informational Quality Labeling on Social Media: In Defense of a Social Epistemology Strategy," *Yale Journal of Law & Technology: Special Issue: Social Media Governance* 23 (June 2021): https://ssrn.com/abstract=3858906.

[66] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi, "Mitigating Political Bias in Language Models Through Reinforced Calibration," In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17 (April 2021): 14857-14866. https://doi.org/10.48550/arXiv.2104.14795.

[67] Kreps and Kriner, "How AI Threatens Democracy."

[68] Elliot G. Morris, *Strength in Numbers: How Polls Work and Why We Need Them* (New York: W. W. Norton & Company, 2022).

One example brought to light by a team of researchers is the outbreak of the war in Ukraine in 2022.[69] Tests using the ChatGPT 3.5 model indicated that while the AI's outputs were fairly accurate on many issues well represented in the model training data, on this emerging issue the model suggested lower levels of U.S. public support for Ukraine than was actually the case in reality. Moreover, the model did not expect that more liberal persons would support U.S. involvement in Ukraine, as they generally have in reality. Further, the model struggled with predicting the correct views of various demographic groups, such as those categorized by age, race, and gender. Model failure to produce accurate data pictures along demographic lines has been found across a wide variety of social and political issues.[70]

Contacting humans in order to gather opinion-related information in a randomized or stratified pattern has only grown more challenging in recent years, given the rise of mobile phones and the difficulty of reaching random samples of people.[71] For years now, this trend has been driving innovation in the polling and survey field, as marketers, political operatives, and a variety of business interests remain hungry for more insights into public preferences and shifts in opinion. The uses of AI models to create, in effect, synthetic citizen respondents will undoubtedly grow more popular as a means of gaining new insights.[72] Some research has suggested that LLMs may be particularly helpful in simulating human responses in sociotechnical systems – for example, the responses of publics to new kinds of algorithms in social media.[73]

However, it is not hard to anticipate the possibilities for recursion in this domain. For example, AI polls might drive political trends: Imagine that a series of AI public opinion simulations are blended with actual polling data, leading to a release of a report that gets wide media

---

[69] Nathan E. Sanders, Alex Ulinich, and Bruce Schneier, "Demonstrations of the Potential of AI-Based Political Issue Polling," *Harvard Data Science Review*, 5(4) (October 2023): doi:10.1162/99608f92.1d3cf75d.

[70] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto, "Whose Opinions do Language Models Reflect?" In *International Conference on Machine Learning* (March 2023): 29971-30004, https://proceedings.mlr.press/v202/santurkar23a.html.

[71] Adam J. Berinsky, "Measuring Public Opinion with Surveys," *Annual Review of Political Science* 20 (May 2017): 309-329, https://doi.org/10.1146/annurev-polisci-101513-113724.

[72] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis* 31, no. 3 (February 2023): 337-51, https://doi.org/10.1017/pan.2023.2.; Jared Council and John McCormick, "Artificial Intelligence Shows Potential to Gauge Voter Sentiment," *The Wall Street Journal*, Nov. 6, 2020, https://www.wsj.com/articles/artificial-intelligence-shows-potential-to-gauge-voter-sentiment-11604704009.

[73] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail, "Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms," *arXiv Preprint* (October 2023): arXiv:2310.05984v1.

attention. A new political challenger candidate – for whatever reasons – seems to be getting little traction in the "polls." That challenger then receives less media attention, driving greater human voter interest in the incumbent. The feedback loop then takes on a life of its own. Such feedback loops are risks across all manner of issue areas, where human preferences may not yet be revealed and dynamic interactions among issues, candidates, and voters may activate tacit knowledge and values that are unpredictable. As with journalism and social media, there are significant epistemic risks to the democratic system in the area of polling. It is another area where cognitive resources and human deliberation may need protection to achieve true alignment and ethical AI goals.

Discussion

Benjamin Franklin's famous quip after the U.S. Constitutional Convention that the delegates had created a "republic … if you can keep it" is perhaps apt as we consider the newest dangers to keeping alive democratic traditions and mechanisms.[74] If good intentions and wise AI governance models prevail, AI technologies may indeed allow greater participation, more deliberation, and greater inputs to democratic systems; AI models may end up being attentive to principles such as beneficence and justice. Technologists speak of increasing the "bitrate of preference communication" within democratic systems.[75] But some dangers threatening human-centered democracy lurk perhaps not in the guise of AI supremacy, killer robot-fueled existential risk, and takeover of humanity but in the slow, grinding effects of AI models mediating what people believe to be true and worthy of attention. The dangers to autonomy lie in feedback loops relating to public knowledge, and epistemic risk will remain ever-present.

For journalism, the danger is that AI, conditioned on past data and inattentive to emerging tacit knowledge and the continuing possibility of unrevealed preferences, continues to do agenda-setting, framing of narratives, and selection of information based on an epistemically anachronistic model. In the domain of social media, the intrusion of AI-powered moderators may erase vital space for human deliberation; and such actions by chatmods may be subject to feedback loops and epistemic risk given the fast-changing environment of human meaning and intention that unfolds on social media. Finally, polling using AI-driven simulations has the

[74] Julie Miller, "'A republic if you can keep it': Elizabeth Willing Powel, Benjamin Franklin, and the James McHenry Journal," Library of Congress Blog, January 6, 2022, https://blogs.loc.gov/manuscripts/2022/01/a-republic-if-you-can-keep-it-elizabeth-willing-powel-benjamin-franklin-and-the-james-mchenry-journal/.

[75] Keiran Harris and Robert Wiblin, host, "Tantum Collins on what he's learned as an AI policy insider at the White House, DeepMind and elsewhere," 80,000 Hours (podcast). October 12, 2023, https://80000hours.org/podcast/episodes/tantum-collins-ai-policy-insider/.

potential to warp the information ecosystem, altering human preferences in democratic space and creating recursive spirals.

These three example areas are important, but this is far from a comprehensive assessment of the epistemic risk problem. For example, research could be extended to other domains, such as online search and discovery. Increasingly, search engines such as Google are consolidating knowledge into single panels of information, making broad browsing of websites and links perhaps much less likely.[76] At the same time, generative AI tools such as ChatGPT, Gemini, and Claude synthesize information from the past web and provide answers to questions using algorithms that leverage probabilistic methods. These generative AI shortcuts may be efficient for citizens, but in short-circuiting traditional processes of browsing, discovery, deliberation, and reasoning they make humans ever-more reliant on the selection and summarization capacities of AI models.

A general framework for attending to this problem of epistemic risk must revolve around the epistemic humility of AI models in the context of democracy and political life and the creation of social norms that emphasize the incompleteness of models and their judgments. There are emerging ideas and frameworks that could make mitigating this risk possible. For example, the guiding idea of "hybrid collective intelligence" attempts to merge the various currents of technocentric optimism, human-centered caution, and enthusiasm over collective intelligence mechanisms and to demand a further set of principles that include continuous oversight of models and explainability. Indeed, emphasizing the normative imperative for explainability – and related concepts such as interpretability and observability – may be a key to attacking the epistemic risk problem. Constant attention to explainability will help to highlight limits and incompleteness, serving as an antidote to the lock-in AI models may create. That said, merely trying to democratize these models by adding more human deliberation around the training and governance of AI models will not be sufficient, as the capacity to ensure democratic values will also rest with large institutions and states.[77] Indeed, there is a great deal more to be said about the role of government and public policy in helping to address epistemic risk, as well as the dangers of government using AI-driven systems in this regard.

It could very well be that, in domains such as autonomous vehicles, logistics, manufacturing, and a wide variety of other areas, advanced AI systems will be vastly superior in terms of planning and execution. Their formidable intelligence will become part of the unquestioned,

[76] Olaf Kopp, "How Google creates knowledge panels," *Seach Engine Land,* June 27, 2022, https://searchengineland.com/how-google-creates-knowledge-panels-386025.
[77] Johannes Himmelreich, "Against 'Democratizing AI,'" *AI & SOCIETY* 38, no. 4 (January 2023): 1333-1346, https://doi.org/10.1007/s00146-021-01357-z.

everyday infrastructure of human life. Yet in the realm of democratic and political life, there will be no clean models to build that are free of potential system error, as models will always lack previously unrevealed preferences, tacit knowledge, and important emergent judgments and values that are produced through subjective interactions – human cognitive experiences with the external world driven by emotion, intuition, and imagination.

Any discussion of AI, public knowledge, and democracy must grapple with the wide variation in information environments across the world. Epistemic risks may be more acute for developing countries and systems that lack sufficient information and knowledge-producing institutions that could surface new inputs to challenge epistemically ossified or manipulated AI systems.[78] Further, lower-resourced countries may lack the political and standards-setting bodies that are necessary for healthy, fair, and just information environments.[79] Authoritarian lock-in leveraging AI is not hard to imagine, given the combination of potential population surveillance by AI systems and the epistemic hardening that is possible as AI, untethered from ethical alignment constraints, potentially takes over public knowledge-producing systems. On the other hand, large companies headquartered in the West are likely to own and deploy advanced AI models, making them less accountable in developing countries because of legal and jurisdictional challenges. More analysis of different forms of governance and their effectiveness in governing AI will be required in the coming years.[80]

Achieving ethical alignment for AI technologies and engaging in smart mechanism design will require technical innovation and careful training of models. However, such technical work is necessary but not sufficient. Some degree of epistemic modesty must be factored/programmed into the design of AI models as they bear on democratic life and deliberation. There must be space to create and preserve human cognitive resources that are not substantially shaped by AI-mediating technologies. If not, the era of AI threatens to become a giant exercise of recursion for democracies. AI models will align with themselves; mechanisms for democratic deliberation will channel preferences that were already shaped by AI-driven public knowledge production in areas such as news, social media, and public opinion. The danger of epistemic risk presents a profound problem in the coming AI era, one requiring constant attention to issues of public knowledge and human autonomy within democratic systems.

---

[78] Fernando Filgueiras, "The politics of AI: Democracy and authoritarianism in developing countries," *Journal of Information Technology & Politics* 19, no. 4 (February 2022): 449-464, https://doi.org/10.1080/19331681.2021.2016543.

[79] Jack Snyder, *Human Rights for Pragmatists: Social Power in Modern Times* IPrinceton, NJ: Princeton University Press, 2022), pp. 148-151.

[80] Michael Veale, Kira Matus, and Robert Gorwa, "AI and Global Governance: Modalities, Rationales, Tensions," *Annual Review of Law and Social Science*19 (June 2023): 255-275, https://doi.org/10.1146/annurev-lawsocsci-020223-040749.