3

Public Opinion in the Age of AI

Joshua D. Clinton, Soubhik Barari, Ethan Busby, Trent D. Buskirk,
Ray Duch, Anna-Carolina Haensch, D. Sunshine Hillygus,
Courtney Kennedy, Kevin Munger, Doug Rivers, Sean Westwood

Abstract: The study of public opinion in political science has always occupied an uneasy space between theory and measurement, referring simultaneously to a latent set of beliefs held by members of a political community and to the responses elicited through imperfect, context-sensitive instruments. The integration of artificial intelligence (AI) into public opinion research makes this tension impossible to ignore. AI offers genuine methodological promise: tools that are faster, cheaper, and more flexible than anything previously available to survey researchers. But the effects of AI also extend further than any prior innovation, in ways that can blur the distinction between measuring and manufacturing public opinion. In doing so, AI brings into sharp relief questions the field has long grappled with: What exactly is public opinion, and what counts as evidence of it? Whether AI enriches or distorts the study of public opinion depends on choices that are as much theoretical as technical; but, taken to the limit, the wholesale substitution of AI-generated responses for human voices may risk a form of epistemic collapse in which public opinion research loses its anchor to the people it purports to represent.

In an early effort to fuse political science with computation, the Simulmatics Corporation promised, in 1960, to infer public opinion by feeding survey data and voter files into a mainframe and asking the machine what "the public" wanted (Pool et al., 1964; Lepore 2020). Nearly two decades later, Douglas Adams's fictional supercomputer "Deep Thought" offered the opposite fantasy: a superintelligence that could answer the Ultimate Question of Life, the Universe, and Everything with apparent authority and precision, leaving humans uncertain what the answer actually meant (Adams, 1980). Today's large language models narrow the distance between these two visions. Trained on vast corpora of human expression – including past surveys, news, and social media – LLMs both generate answers and simulate the publics that might have produced them. In doing so, they often blur the line between measuring and manufacturing public opinion.

Like earlier technological shifts – from mass polling to television to social media – AI reshapes both how opinions are formed and how they are measured. AI offers genuine methodological promise: tools that are faster, cheaper, and more flexible than anything previously available to survey researchers. It also expands what might possibly count as evidence of public opinion, opening up text, audio, video, and social media to analysis at unprecedented scale. It transforms every stage of the survey process, from instrument design to interviewing to post-collection analysis. And it can even bypass the survey process altogether, generating synthetic respondents in place of the humans that survey research exists to reach, a possibility that is as conceptually unsettling as it is technically impressive.

The challenges posed by AI to public opinion are perhaps most visible in the emergence of "silicon respondents" – AI-generated personas prompted to answer survey questions as members of a target population. Silicon respondents offer a seductive promise: on-demand data, cheaper access to hard-to-reach groups, and freedom from the practical constraints that have made survey research increasingly costly. The risk, however, is not simply technical error, but the replacement, distortion, or outright manufacture of the human voices that public opinion research exists to represent. Using AI in survey research often implicitly embeds assumptions about what public opinion is and how it can be observed. AI does not merely change how we measure public opinion. It compels us to say what we mean by it.

This chapter takes that compulsion seriously. We examine how AI reshapes public opinion as a concept, as a social reality, and as a set of measurement practices on which scholars and political institutions depend. The deepest question is not whether the technology works, but whether its use preserves the epistemic and democratic commitments that motivate the study of public opinion in the first place. At stake is not merely methodological accuracy, but the risk that AI-mediated representations of public opinion gradually detach from the publics they are meant to describe by amplifying dominant voices, obscuring the hardest-to-reach, and substituting plausible-seeming but misleading outputs for the human voices they purport to represent. In the limit, the result is not merely measurement error but a form of epistemic collapse in which public opinion research loses its anchor to the publics it was built to elevate.

I. The Concept of "Public Opinion"

It is difficult to overstate the importance of public opinion to governing and governments. Even authoritarian regimes must attend to public sentiment, if only to avoid provoking instability or revolt. As V. O. Key famously observed, "Governments must concern themselves with the opinions of their citizens, if only to provide a basis for repression of disaffection" (Key, 1965). In representative democracies, public opinion occupies an even more central role, shaping debates over political

responsiveness, accountability, and legitimacy. Longstanding normative disagreements persist over whether elected officials should exercise independent judgment or defer to constituent preferences (Burke, 1774), and some have argued that responsiveness to public opinion represents the highest form of democratic government (Bryce, 1888).

Despite its centrality, "public opinion" is a remarkably elusive concept. Scholars have defined it in multiple, often incompatible ways: as the aggregation of individual preferences (Key, 1961), the considered judgments of an informed and deliberative public (Fishkin, 1991), the expression of organized social groups (Blumer, 1948), or a mediated construction shaped by elites, institutions, and information environments (Lippmann, 1922). Even abstracting from problems of measurement, classical results in social choice theory demonstrate that aggregating individual preferences may yield indeterminate or unstable collective outcomes, even when individual preferences are internally consistent (Arrow, 1951). At the individual level, many citizens lack well-formed or stable political attitudes, instead offering responses that reflect the wording of a question, contextual cues, or sampled considerations, rather than durable beliefs (Converse, 1964; Sniderman et al., 1991; Zaller, 1992). The conceptual foundations of public opinion are therefore unsettled even before empirical measurement enters the picture.

Efforts to measure public opinion introduce additional complications. Survey instruments do not merely record pre-existing attitudes; they actively participate in constructing public opinion by defining who counts as the public, what counts as an opinion, and how attitudes are elicited (Herbst, 1993). Survey responses may reflect momentary considerations rather than underlying preferences, and they are sensitive to question wording, order, and context (Schuman and Presser, 1981). Patterns of unit and item nonresponse further complicate inference when those who decline to participate differ systematically from those who do (Groves and Couper, 1998; Berinsky, 2004; Kennedy et al., 2018; Clinton et al., 2021). Even under ideal conditions, surveys provide an imperfect and contingent window into public sentiment rather than a transparent record of latent attitudes. These complications are not merely technical: They raise a more fundamental question about what public opinion research is actually trying to measure. Is the goal to characterize observed sample responses, to recover the latent opinions those responses may imperfectly reflect, or to make inferences about the broader population from which respondents are drawn? Although we typically assume the goal is the latter, each target raises different challenges and implications for how AI tools should be understood and evaluated.

Technological Change and Public Opinion

Technological change has long reshaped both the formation of public opinion and the methods used to observe it. The rise of scientific polling in the early twentieth century did not simply provide a new measurement tool; it helped construct the very idea of "the public" as a coherent and measurable political entity (Gallup, 1940; Igo, 2007). Subsequent innovations in mass communication, from radio to television, altered how citizens encountered political information by reshaping persuasion, agenda-setting, and collective attention (Lazarsfeld et al., 1944; Iyengar and Kinder, 1987). Later transitions in communication infrastructure from landlines to cell phones and internet-based surveys further destabilized sampling frames and modes of inference in ways that repeatedly forced scholars to reconsider how opinions were formed, whose views were visible, and how representative any given measure of public sentiment could be (Dillman 2000; Groves et al. 2009).

Recent advances in AI extend this trajectory by making it possible to systematically analyze a far wider range of human expression than was previously feasible. As AI systems become capable of processing text, audio, video, and images, sources such as social media posts, videos, and qualitative interviews can be analyzed using methods of statistical analysis and summary long associated with survey research (e.g., Alshaabi, 2021). Because these systems are trained on vast corpora of human language and behavior, the patterns they extract reflect regularities in human expression at scale (Caliskan et al., 2017; Cheng et al., 2023; Hofmann, 2024). In this sense, AI blurs the boundary between survey-based measures of public opinion and other forms of expression that were previously ignored due to limits of scalability, standardization, analytic tractability, and consent to being measured.

Social media illustrates both the promise and the limits of this shift. Interest in "social media revolutions," such as the Arab Spring, spurred research on how online expression could disrupt pluralistic ignorance and destabilize unpopular regimes. In stable democracies, scholars turned to sophisticated poststratification methods – statistical techniques that reweight samples to better match the population of interest – to address the non-representativeness of social media users and extract implied public opinion from platforms such as Twitter (now called X). AI plausibly re-invigorates this line of research by relaxing earlier constraints on scale and flexibility, particularly as political expression increasingly takes visual and multimodal forms (Munger, 2024). At the same time, greater analytic capacity does not overcome enduring conceptual limitations. Whether AI-analyzed expressions constitute evidence of public opinion depends entirely on which target is in view. If the goal is to recover latent opinion rather than observed survey responses, digital traces may be informative and people's unelicited expressions may reflect their beliefs more authentically than responses to survey instruments. But the population problem looms even larger here than in survey research: It is far harder to assess how representative any given corpus of expression is, whose voices are captured, and how the patterns which AI extracts should be mapped onto any identifiable public (see, for example, Baack, 2024; Jungherr et al., 2016). For these reasons, surveys remain indispensable even as AI expands what may count as evidence of public opinion.

Arguably the most fundamental shift associated with social media is the way it directly encodes information about public opinion alongside media content (Munger, 2020). Because social media is inherently social, users are continually exposed to not only political information but to cues about what others think. These cues shape beliefs about the broader cultural network in which individuals are embedded (Munger, 2024), expand awareness of the political views held by acquaintances (Settle, 2018), and contribute to the diffusion of political meaning across an increasingly wide range of identities, preferences, and consumer goods – the "oil-spill" model of politicization (DellaPosta, 2020).

If social media reshaped public opinion by altering the flow of social information, AI represents a potentially more profound shift in how individuals encounter, interpret, and generate politically relevant content. Social media primarily affects what people believe *others* think, but AI systems intervene more directly in processes of information seeking, judgment formation, and expression. Rather than merely mediating or reporting information, AI tools can generate content, personalize it at scale, and influence how information is weighted and surfaced by downstream systems. AI-generated bots may shape public opinion through targeted content or by affecting the signals that guide platform algorithms (Schroeder et al., 2026), and systems optimized for affirmation may reinforce inferred beliefs rather than promote updating or exposure to competing perspectives (Bisbee et al., 2026). Recent experiments confirm that AI chatbots can durably reduce conspiracy beliefs and shift voter preferences by substantial margins across multiple countries (Costello et al.,

2024; Hackenburg and Margetts, 2024; Lin et al., 2025); and, audits suggest that model outputs tend to lean more liberal and cosmopolitan than the average voter (Santurkar et al., 2023; Westwood et al., 2025; Lyman et al. 2025). While not the focus of our chapter, understanding how public opinion itself is shaped by people's encounters and usage of AI in ordinary contexts is a topic of increasing importance.

II. The Effect(s) of AI on Survey-Based Measures of Public Opinion

Given the conceptual stakes outlined above, how AI is actually being integrated into survey practices matters enormously, and recent work has begun the task of categorizing and organizing AI's effects across the survey lifecycle – from design and administration to analysis and reporting (Rothschild et al., 2025). As AI applications in survey methodology continue to expand in ways that make any summary necessarily provisional, it is useful to highlight the larger conceptual issues about the nature of public opinion that may be raised by their use. Depending on one's perspective, the use of AI in survey research can be understood in at least three different ways, with differing implications for what is meant by public opinion:

- as a technical tool that makes survey research more efficient by expanding the types of questions we can ask (including open-ended items) and enabling automated pretesting and power calculations;
- as an inferential tool that helps address item and unit nonresponse through increasingly sophisticated, but often opaque, imputation algorithms; and,
- as a replacement for surveys altogether as an "on demand" public opinion system in which researchers query the corpus directly rather than fielding a survey to humans.

Whereas the first use largely avoids conceptual claims about the epistemological status of AI outputs, imputation-based uses require treating AI-generated responses as meaningfully equivalent to human survey responses. The third use case raises the most substantial conceptual issues, as it remains unclear whether silicon respondents are intended to proxy human survey responses, with their attendant artifacts, or intended to represent estimated latent preferences independent of the survey process. All three uses typically assume that population-level inference is the ultimate goal. But the uses differ considerably in how directly they engage the conceptual ambiguity that population-level inference entails. And they differ in what assumptions they require researchers to make about the relationship between observed responses, latent opinion, and population distribution.

The meaning of AI responses and respondents is increasingly important even for those seeking to avoid silicon respondents altogether because of the potential contamination of online nonprobability surveys by either silicon respondents or responses resulting from a human using AI to generate responses (Westwood, 2025). The stakes are extremely high; in earlier eras, the principal concern was the mismeasurement of human opinion; but the AI era offers the more unsettling possibility that researchers may be unable to know whether the responses they are analyzing are created by humans or non-humans. AI is not simply another methodological tool or technical complication; its use compels a reconsideration of what the public is, how its views are formed, and what it means to measure public opinion. These conceptual questions become concrete in the practical decisions researchers now face about how, and at what stage, to incorporate AI into survey research.

Turning to the ways in which AI is used in survey methodology, Buskirk and colleagues (2025a) explore the uses of large language models within three broad phases of the survey research

process: tasks prior to data collection, data collection itself, and post data-collection activities (see also Rothschild et al 2026). Most uses currently involve data collection and post data-collection phases. Figure 1 reports examples that include: *coding open-ended responses* (Mellon et al., 2024; Rytting et al., 2023; Singh and Kumar, 2025; von der Heyde et al. 2025a), *interviewing* (Barari et al. 2025; Lang and Eskenazi 2025; Wuttke et al. 2025; Xiao et al. 2020), *pretesting and survey development* (Buskirk et al., 2025b Adhikari et al., 2025; Tao et al., 2024; Yun et al., 2024), and *analyzing responses* (Bodin, 2024 and Huang et al., 2024).

When thinking about the many ways in which AI may affect survey research, it is useful to consider the Total Survey Error framework, a systematic accounting of all potential sources of error in survey research – from sampling and recruitment through measurement and response. The Total Survey Error framework helps organize how various AI interventions may affect the error and risks associated with each step. Although AI may substantially mitigate some forms of survey error, systematic evaluations of its effects in other applications are more limited, and the framework emphasizes the importance of considering whether and how AI-based approaches affect data quality or validity relative to traditional methods. Existing work has yet to fully integrate AI into the Total Survey Error framework, but applying that framework going forward offers a structured way to evaluate how AI may both reduce and generate distinct forms of survey error.

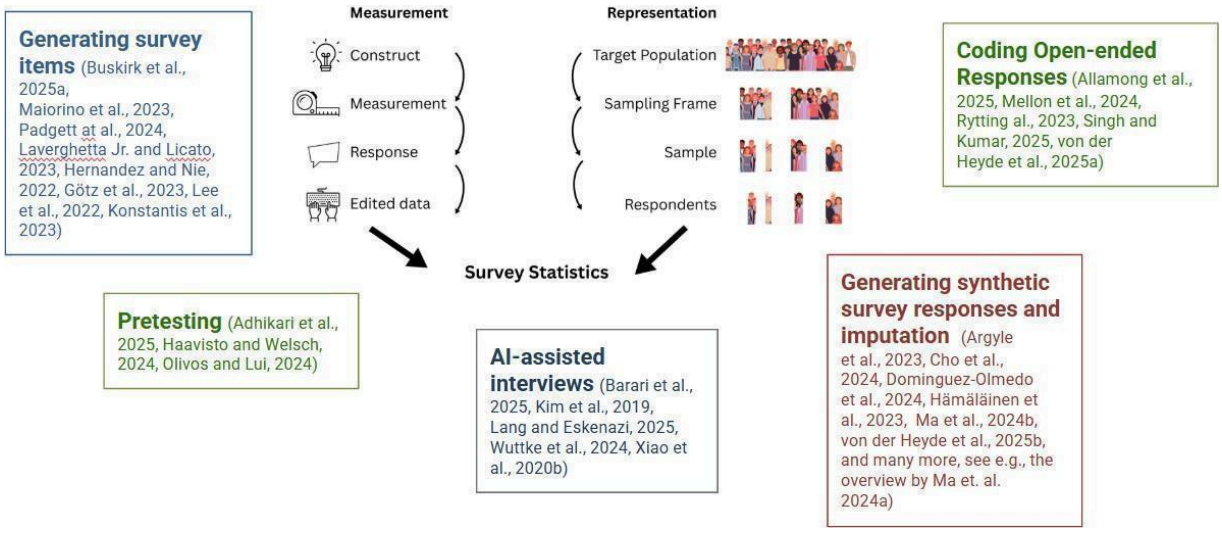

Figure 1 Selected uses of AI in survey research. Illustration shows some ways AI is currently used in survey research: Generating survey items, pre-testing, AI-assisted interviews, generating synthetic survey responses and imputation, and coding open-ended responses. Source: Own depiction, by Anna-Carolina Haensch of the Total Survey Error, adapting Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Pre-Data Collection

Recent work has explored several ways in which AI can be incorporated into the tasks required prior to the collection of data, including questionnaire writing, item generation, translation, sampling designs, and recruitment. Such applications fall largely within the first category of AI usage – AI as a technical tool that enhances efficiency without fundamentally altering what is being measured. But even here, the boundary is not always clean and more interpretive uses of AI in pretesting can begin to embed assumptions about respondent cognition that carry inferential weight.

A growing line of research examines the use of LLMs for generating surveys at both the item and questionnaire level (e.g., Behrend and Landers, 2025). Empirical findings suggest that LLMs can produce relevant and comprehensible questions with response options that align closely with the intended topic (Maiorino et al., 2023; Padgett et al., 2024). However, question quality depends heavily on prompt design and input structure. Models tend to exhibit consistent stylistic regularities, such as preferences for binary or five-point Likert response formats, difficulties with causal precision, and challenges maintaining coherent ordering across longer instruments (Padgett et al., 2024; Buskirk et al., 2025b; Fuchs et al., 2025). Absent additional constraints, generated items are often written at relatively high reading levels, underscoring the need for independent evaluation using established survey quality metrics (Buskirk et al., 2025c, Olson and Buskirk, 2025). Overall, systematic evidence remains limited as to the measurement quality and construct validity of AI-generated survey items.

A comparatively less controversial application of AI in the pretesting stage involves checking questionnaire logic, flow, and programming. AI systems can help identify broken skip patterns, inconsistent branching logic, missing response options, and mismatches between question wording and programmed responses. Because these applications focus on mechanical properties of the instrument, they resemble existing automated testing practices and pose few conceptual risks.

More expansive and interpretive uses of AI in pretesting that involve the simulation of respondent understanding or generating responses from hypothetical personas introduce greater inferential risks. Studies show that LLMs can flag potential design flaws or provide preliminary feedback prior to human pretesting, but performance depends on prompt design and the availability of best-practice examples (Tao et al., 2024; Olivos and Liu, 2024). When used to generate responses for exploratory analysis or theory development, such approaches raise concerns about representativeness, bias, and model-induced agreement, particularly when outputs are treated as informative about real populations rather than as heuristic inputs (Anthis et al., 2025; Park et al., 2024a).

A distinct but related application of AI involves translation and cultural adaptation of survey instruments. Some studies find that AI-assisted translation can achieve semantic quality comparable to professional translators and help flag issues related to ambiguity, formality, and cultural specificity (Haavisto and Welsch 2024; Metheney and Yehle, 2024; Adhikari et al., 2025). However, evidence outside Western, educated, industrialized, rich, and democratic contexts remains limited and some work finds mistranslation of gender-neutral language, the imposition of culturally specific values, and sensitivity to prompt language and model choice (Ghosh and Caliskan, 2023, Tao et al., 2024; Lee et al., 2025).

Data Collection

AI can also be incorporated in the data collection phase of survey research, most prominently through AI-assisted interviewing. These approaches use AI models to conduct interactive, conversation-based interviews that tailor question wording, sequencing, and probing to respondents. This can range from limited adaptation within otherwise fixed surveys to fully interactive chatbot-style interviews that generate personalized question sequences. Although labeled variously as AI adaptive interviewing, conversational AI, or chatbot interviewing, we refer to these approaches collectively as AI-assisted interviewing. AI-assisted interviewing sits at the boundary between the first and second uses of AI identified above. When adaptation is limited to clarifying fixed items, it functions primarily as a technical efficiency tool that leaves the inferential target largely intact. But as personalization deepens (e.g., tailoring probes, reordering questions, or generating follow-ups), the instrument itself becomes partially AI-constructed. This raises the question of whether the resulting responses reflect the respondent's latent opinion more faithfully, or whether those responses are shaped by the interaction with AI in ways that blur the line between elicitation and construction.

AI-assisted interviewing offers several potential methodological benefits, including the ability to clarify concepts to the respondent or interviewer while interviewing, real-time analysis of open-ended responses, and adaptive branching (Barari et al., 2025; Lang and Eskenazi, 2025). Personalization may improve probing of core issue positions or time-sensitive, high-salience topics (Velez and Liu, 2025; Velez, 2025), and some studies report that AI-assisted interviews elicit more informative and contextually relevant responses than traditional web surveys (Barari et al., 2025; Xiao et al., 2020). The use of AI may also reduce interviewer effects (West and Blom, 2016) and social desirability bias (Krumpal, 2011), increase engagement (Yun et al., 2024; Lang and Eskenazi, 2025), and offer advantages in scalability, cost, and operational efficiency (Wuttke et al., 2025). However, these benefits are design- and mode-dependent: Increased interactivity can also raise cognitive burden and break-off rates for respondents, perhaps particularly in mobile contexts (Barari et al., 2025).

AI-assisted interviewing can also introduce new sources of error. Respondents' comfort with AI interviewers varies and can affect participation, disclosure, and data quality (Tirumala et al., 2025). Some respondents disclose more sensitive information to AI interviewers due to perceived anonymity (Xiao et al., 2020), raising ethical and privacy concerns. A further challenge is balancing personalization with standardization: Because AI-assisted interviews may vary wording and probes across respondents, they may reduce comparability relative to fixed instruments. In addition, the prompting and training of AI-based interviewers may embed researcher assumptions in ways that affect the questions being asked and the responses being given. It may be difficult to determine whether the gains from personalization reflect respondents providing more authentic and unfiltered information or the influence of researcher-defined prompts and interactional norms. This ambiguity maps directly onto the distinction between observed responses and latent opinion: If AI-assisted interviewing affects *what respondents say*, researchers must determine whether it does so by reducing measurement error by obtaining responses that better reflect respondents' actual latent opinions or by introducing a new source of error.

A closely related concern is the growing prevalence of AI-assisted responding, in which respondents themselves use AI tools when completing online surveys. Early evidence suggests this occurs at substantial rates among experienced nonprobability respondents and poses a serious threat to data integrity (Martherus et al., 2025; Pinzón et al., 2025; Traylor, 2025). For example, LLM-generated responses to open-ended items tend to be more homogeneous and more positive than human-written

responses, particularly for sensitive social groups. This potentially distorts downstream representations of public opinion (Zhang et al., 2025).

This problem is likely to intensify as autonomous AI agents capable of completing surveys without human input become more prevalent. Unlike the intentional use of silicon respondents discussed below, AI-agent participation introduces silicon responses into surveys without researchers' awareness or consent, raising inferential concerns even for scholars who reject their use. Such responses may be difficult to detect: Westwood (2025) and Martherus et al. (2025) show that AI agents, including Operator, can evade standard detection methods and even self-identify as human.

Although no reliable or definitive detectors of AI-assisted responding currently exist, contamination of online survey data is likely to increase as language models improve. Detection strategies based on statistical, linguistic, or behavioral regularities are inherently fragile in adversarial settings, as AI can both learn detection rules and to circumvent them. Even so, researchers have proposed partial tools that may help flag suspicious responses, including metadata signals, paradata anomalies, and semantic diagnostics (Pinzón et al., 2025; Sepulvado, 2025). Open-ended responses can also serve as useful diagnostics of response quality, though overly demanding prompts may inadvertently incentivize AI assistance among respondents (Zhang et al., 2025).

As AI-assisted responding is likely to remain a persistent feature of the survey environment, the methodological challenge shifts from elimination to management. This includes determining which forms of AI assistance, if any, should be treated as acceptable and recognizing that some uses (e.g., translation assistance for respondents with limited English proficiency) may not be inherently problematic. The need to define safeguards and exclusion criteria *ex ante* is therefore central, particularly given that freshly recruited probability-based samples may face fewer risks than opt-in panels with extensive respondent histories.

These considerations suggest that researchers will require layered approaches combining sampling design, providing survey instructions, monitoring metadata and paradata, and ensuring transparent preregistration of data-quality decisions. More fundamentally, however, the emergence of AI agents capable of autonomously completing surveys unsettles the long-standing assumption that responses can be presumed human. As the boundary between human and automated responding becomes less reliably detectable, how samples are constructed, monitored, and interpreted becomes central to preserving data quality and inferential validity (Westwood, 2025). AI may reshape survey data collection in productive ways, but its responsible integration depends on deliberate design choices and sustained human oversight.

Post-Data Collection

AI affects the post-data collection stages of survey research by expanding the tools available for processing, imputation, and interpretation, while introducing new challenges for bias, replicability, and inference (Buskirk et al., 2025a). Among the least controversial uses is assistance with data analysis and reporting. AI tools can support exploratory analysis, recommend analytic strategies conditional on hypotheses and variable structures, and automate routine tasks such as code generation, estimation, visualization, and the drafting of toplines without altering underlying data (Jansen et al., 2023; Chintakunta et al., 2025). A key distinction separates lower-risk uses – where AI generates code that researchers inspect and execute within standard workflows – from higher-risk uses in which AI systems directly perform analytic tasks, potentially obscuring decisions, reducing transparency, and increasing the risk of undetected errors or hallucinated inferences.

Another increasingly prevalent and relatively low-risk application involves coding and analyzing open-ended survey responses (Than et al., 2025). Open-ended responses have traditionally required costly and potentially inconsistent human coding due to variation in interpretation and attention (Reja et al., 2003), leading researchers to favor forced-choice items despite the limitations of the responses. AI-based text analysis offers a scalable and replicable alternative that can reduce post-data collection measurement error while preserving respondent-generated content. In addition to classification, LLMs can assist with preprocessing tasks, such as spelling correction, text normalization, and basic data cleaning (Allamong et al., 2025). Building on earlier supervised and unsupervised approaches (Barberá et al., 2021), recent evaluations generally find strong performance for AI-based classification in well-defined tasks. Mellon et al. (2024) show that top-performing LLMs can nearly match human coders and outperform supervised models trained on 1,000 labeled cases, with comparable findings across political science contexts (Rytting et al., 2023).

Performance, however, depends on implementation choices. Classification accuracy varies with prompt design, category definitions, and model configuration (Shi et al., 2024), and domain-specific biases can emerge, such as lower accuracy for abortion-related responses expressing pro-life positions. Errors may also vary systematically across respondents and contexts. Ashwin et al. (2025) show that, in interviews with Rohingya refugees, LLMs overpredict sparse codes and produce false positives correlated with respondent characteristics, such as gender, education, and refugee status. Multilingual and culturally complex data require additional adaptation and careful attention to rare categories (Singh and Kumar, 2025, von der Heyde et al., 2025).

A conceptually distinct set of post-data collection applications involves using AI to impute missing survey data due to item nonresponse, attrition, or skipped questions. Unlike the coding and analysis tools discussed above, which operate on observed responses, imputation implicitly treats model-generated values as estimates of what respondents would have said – a substantively stronger claim whose assumptions are correspondingly harder to evaluate. It also raises an immediate question about the inferential target: Should imputed values approximate the survey answer a respondent would have given, complete with the satisficing, framing effects, and social desirability bias that characterize observed data? Or should they attempt to recover the latent opinion those responses imperfectly reflect? The two targets require different models, different training data, and different standards of validation, yet the distinction is rarely made explicit in practice. Imputation differs from the use of silicon respondents discussed below because imputation estimates responses conditional on observed human data rather than generating them de novo. Kim and Lee (2023) identify three analytically distinct imputation problems: (1) conventional imputation to generate responses for asked but unanswered questions; (2) retrodiction, or the reconstruction of responses to past questions based on current responses; and (3) unasked opinion prediction, which attempts to infer attitudes on never-asked questions. They show that fine-tuned LLMs trained on representative survey data can perform well for the first two tasks, reconstructing known opinion shifts while maintaining stability during exogenous shocks. But performance declines when predicting unasked opinion and varies considerably across demographic groups.

As with all imputation, using AI to impute survey responses raises important ethical and methodological concerns, particularly when imputing sensitive attitudes or behaviors. Unlike traditional methods, however, the effects of AI-based imputation approaches are harder to evaluate because their assumptions are difficult to specify and interrogate. Conventional imputation models make functional forms, conditioning variables, and missingness mechanisms explicit, whereas AI-based imputations rely on high-dimensional representations learned from opaque training corpora. This opacity increases

the risk that model-generated responses default to typical patterns, overstate precision, mask heterogeneity, and blur the distinction between measured responses and model-based inference. Scholars must therefore distinguish carefully between observed data and AI-derived estimates.

AI's post-data collection uses range from efficiency-enhancing tools for coding and analysis to more consequential forms of inference that reshape how missing data are handled. We currently lack a systematic assessment of how these applications affect Total Survey Error (but see Rothschild et al., 2026), underscoring the need for caution and clearer evaluative standards as AI becomes embedded in post-survey workflows. The most extreme form of this progression in which AI-generated data replace that of human respondents entirely is the subject to which we now turn.

III. Silicon Respondents

Perhaps the most transformational, and controversial, way in which AI reshapes the conceptualization and measurement of public opinion is through the creation of silicon respondents (Argyle et al., 2023; Horton, 2023; Park et al., 2024; Wang et al., 2025). Silicon respondents are often generated by prompting LLMs to adopt a persona defined by demographic, attitudinal, or contextual information – and answer survey questions as that persona to approximate how members of the adopted target population might respond. Unlike classical imputation or synthetic data techniques grounded in explicit probabilistic models such as multiple imputation by chained equations or predictive mean matching (Drechsler et al., 2023), AI-generated responses do not emerge from transparent statistical assumptions about sampling or missingness. Instead, they reflect an opaque combination of prompts, patterns learned from massive but not necessarily representative training corpora, and post-training alignment procedures.

Silicon respondents represent not only a new methodological tool but also a challenge to foundational assumptions about measurement error, response artifacts, and the nature of public opinion itself. Public opinion research has long rested on the premise that surveys capture the expressed attitudes of real people at a particular moment in time, albeit imperfectly and subject to the various sources of error identified by the Total Survey Error framework. By making it possible, in principle, to generate public opinion data without human respondents, silicon respondents force a confrontation with what we mean by public opinion. If the target is survey responses of the observed sample, silicon respondents must reproduce the distribution of responses that human respondents provide, artifacts and all (e.g., satisficing, social desirability bias, question framing effects). If the target is latent respondent opinion stripped of survey-induced error, silicon respondents must recover what people actually believe rather than how they respond when asked survey questions. And if the target is population-level responses or opinion, silicon respondents must additionally address issues of representation – whose views are captured in the training data, which populations are overrepresented or absent, and can any adjustment recover the distribution of opinion in the broader public? These are not merely technical questions. Each target implies a different standard for what counts as a valid benchmark, a different conception of what silicon respondents are being asked to do, and a different basis for evaluating whether they succeed. The use of silicon respondents forces these distinctions to be confronted explicitly in ways that conventional survey research has often been able to leave implicit.

Independent of this conceptual ambiguity is the difficulty of establishing clear evaluative criteria. Conclusions about the ability of silicon respondents to mimic human responses depend heavily on the properties being assessed. Concerns are often framed in terms of algorithmic bias arising from training data (Bender and Friedman, 2018), post-training alignment (Lyman et al., 2025),

or prompt design (Shubham et al., 2024; see also Gallegos et al., 2024; Hu et al., 2025; Westwood et al., 2025), but the validity of silicon respondents depends on the aggregate effects of each. The central question is not whether a model exhibits bias in how it responds to a particular prompt, but whether it can adopt personas corresponding to real populations of interest and generate responses that accurately reflect the joint distribution of human opinions both within and across questions. Scholars have proposed a range of criteria for evaluating silicon respondents (Röttger et al., 2024). Early work emphasized overall correspondence between human and silicon responses (Argyle et al., 2023), while subsequent studies focused on variance and distributional properties (Bisbee et al., 2024) and subgroup-level correspondence within populations (Santurkar, 2023; Kim and Lee, 2023). Different approaches prioritize different evaluative targets and reach divergent conclusions about the usefulness and reliability of silicon respondents.

Some studies find close agreement at an aggregate level, particularly in the distribution of responses and their covariance with traits such as partisanship and ideology (Argyle et al., 2023). Others argue that this apparent similarity is superficial, pointing to substantial noise and systematic inaccuracies in conditional relationships (Santurkar et al., 2023). Still others show that performance varies sharply by context and is highly sensitive to prompting choices (Bisbee et al., 2024; Li and Qi, 2025), or that silicon responses exhibit less variation than comparable human data because they homogenize group differences (Bisbee et al., 2024; Wang et al., 2025).

Common to all evaluations to date is the reliance on algorithmic fidelity – the assumption that silicon responses should mirror human survey responses to some degree (e.g., Argyle et al., 2023; Lee et al., 2024; Amirova et al., 2024; Lyman et al., 2025). This benchmark is intuitively appealing, and silicon respondent performance has improved as AI models have become more capable (e.g., Wang et al., 2025). At the same time, this approach is necessarily limited in at least two respects.

The most fundamental limitation of algorithmic fidelity is that validation necessarily relies on historical or concurrent human survey data. Reproducing past survey distributions is of limited value since those data already exist, and concurrent benchmarks are unavailable precisely where silicon respondents are most appealing: generating estimates for populations or moments where human data cannot be collected. Even strong performance against available benchmarks does not imply fidelity going forward, particularly as events, contexts, and opinions change.

A second limitation concerns what algorithmic fidelity is actually measuring. The standard benchmark treats correspondence with observed human survey responses as the validation criterion. But observed sample responses are not the same as population-level opinion, and the gap between them matters. With human surveys, the sources of non-representativeness are at least partially observable, and researchers can in principle adjust for them through weighting and post-stratification. With silicon respondents, the effective sampling frame is the LLM's training data, which is opaque: It is not possible to know which populations are over- or underrepresented in ways that would allow principled correction. Matching a particular human sample, therefore, does not establish that silicon respondents recover population-level opinion. The problem is compounded if the goal is latent opinion rather than observed survey responses. If what we care about is what people actually believe, independent of the artifacts introduced by the survey process itself, then fidelity to survey responses embeds the very distortions, including satisficing, framing effects, and social desirability bias, that the concept of latent opinion is defined against. In short, algorithmic fidelity is most defensible precisely when it is least needed, when high-quality representative human survey data already exist, and least defensible where silicon respondents are most appealing, when such data are unavailable, unrepresentative, or potentially distorted. Alternative benchmarks assessing correspondence with

population-level behavioral outcomes, such as turnout, election results, or political donations, may be more appropriate in some contexts (e.g., Cerina and Duch, 2023), though these too carry assumptions about what counts as ground truth.

When human survey data are carefully designed, well-fielded, and appropriately weighted, they remain one of the most, if not the only, credible tools for measuring public opinion (Jamieson et al., 2023). In such cases, comparisons between silicon and human responses can be informative about the extent to which AI models can replicate expressions of public opinion as reflected in high-quality surveys. However, the ability to assess silicon responses becomes far more difficult when survey quality is uncertain because close agreement to a poorly measured survey does not establish accuracy, and divergence from a weak benchmark does not necessarily imply inaccuracy. Claims about the fidelity or validity of silicon respondents must, therefore, be interpreted in light of the design quality and measurement properties of the surveys to which they are being compared and should remain anchored to empirical comparison with high-quality human data whenever possible.

In practice, the use of silicon respondents frequently proceeds without resolving this ambiguity, leaving uncertain both what is being measured and whether the use of silicon respondents reflects institutional or economic incentives, rather than a coherent position on what public opinion is and how it ought to be measured. Given this unresolved ambiguity, it is useful to consider the spectrum of ways in which silicon respondents may be used, ordered roughly by the importance of resolving the underlying conceptual question.

Even scholars who are committed to the exclusive study of human respondents cannot fully ignore silicon respondents because understanding their characteristics is increasingly important for assessing the measurement error that AI-assisted respondents may introduce into online surveys. If human respondents rely on AI tools to draft or refine answers, for example, the behavior of those tools will shape observed data. Scholars committed to analyzing exclusively human responses must, therefore, engage with silicon respondents, if only to use them as diagnostic tools for identifying and evaluating emerging sources of survey error.

Beyond this diagnostic role, silicon respondents may also have instrumental value for study design and power analysis. They can be used to explore hypothetical relationships much as agent-based models do – assessing whether planned analyses appear able to detect anticipated effect sizes or determining required sample sizes given plausible effect distributions (e.g., Sarstedt et al., 2024; Dillon et al., 2023; Ashokkumar et al., 2024; Lippert et al., 2024). In these applications, silicon respondents are not obviously inferior to existing practices, but caution is warranted because of possible silicon demand effects, in which AI systems infer experimental hypotheses and generate responses that artificially confirm them (Westwood, 2025).

A more maximalist approach uses silicon respondents to replace some or all human respondents entirely, whether through zero-shot prompting or by fine-tuning models on human opinion data. Fully embracing this approach pushes the study of public opinion to prioritize simulated attitudes over the responses of interviewed humans. Although this position is not the dominant view among public opinion scholars, given concerns about the properties of silicon responses (e.g., Bisbee et al., 2024; Wang et al., 2025), it is notable that leading survey providers are already offering entirely-silicon data for market and consumer research (e.g., https://www.qualtrics.com/edge/). And numerous companies and startups now offer silicon respondents on demand (e.g., https://www.electrictwin.com/, https://askditto.io/, https://www.expectedparrot.com/, and https://www.syntheticusers.com/ to name but a few).

A larger, arguably existential risk arises from the widespread use of silicon respondents. Because the ability to reproduce past survey data is of limited value because those data already exist, the greatest potential value of silicon respondents lies in providing estimates when human data are unavailable. But if new human data are increasingly replaced by responses from silicon respondents, the entire research ecosystem risks model collapse, a phenomenon in which AI systems become progressively worse when trained on their own outputs (Shumailov et al., 2024).

The possibility of model collapse exposes a broader limitation of prevailing validation strategies. A researcher may save resources or improve performance in a single application, yet in the aggregate such choices can erode the informational foundations on which future models depend. Evaluating silicon respondents in isolated settings fails to account for the ways in which individual methodological decisions generate spillover effects across the research ecosystem. This concern extends to alternative validation metrics as well: Behavioral benchmarks, such as election outcomes or donation patterns, may themselves already be embedded in model training data, making it difficult to treat them as independent ground truth. Indeed, silicon respondents may appear most reliable precisely when they are least dangerous, before their widespread adoption reshapes the data ecosystem on which subsequent validation relies. If pollsters, scholars, policymakers, and citizens increasingly substitute faster and cheaper silicon estimates for direct engagement with human respondents, and if each iteration compresses variation or privileges dominant viewpoints, representations of public opinion may gradually detach from the populations they are intended to describe. If so, public opinion research risks a form of epistemic collapse.

An intermediate, hybrid approach combines silicon data with human survey data in an effort to reduce bias under conditions of extreme nonresponse or selective participation (e.g., Berinsky, 2004; Kennedy et al., 2018; Bailey, 2024). Rather than replacing human respondents altogether, hybrid designs use silicon respondents to approximate the views of individuals or groups who exist in the population but do not participate in surveys (Duch et al., 2024, Cerina and Duch, 2023). In principle, such augmentation may improve inference if human-only samples are increasingly unrepresentative.

Yet hybrid approaches rest on strong and often unverifiable assumptions: that the data used to construct silicon respondents are representative of nonrespondents; that online traces or model-generated profiles can be meaningfully translated into survey responses; and that human and silicon data can be integrated in ways that recover population-level attitudes rather than distort them. Because the opinions of nonrespondents are unobserved, there is no stable ground truth against which to evaluate whether augmentation reduces bias or merely substitutes one form of error for another. Whether hybrid designs represent an improvement depends not only on model performance, but on explicit judgments about which assumptions are more defensible and which risks are more tolerable.

Although the use of silicon responses may reduce costs and expand access, particularly for resource-poor teams or hard-to-survey populations, neither advantage ensures validity or reliability. Proprietary frontier models may offer superior performance at the expense of transparency and reproducibility relative to open-weight alternatives, and performance may vary greatly based on pre-training, prompting, and post-model calibration. As a result, nominally similar silicon respondents may reflect meaningfully different assumptions and behaviors across models and implementations, making it difficult to determine whether observed differences arise from substantive features of public opinion, from contingent modeling choices, or from changes in the model systems themselves. Using silicon respondents therefore requires reasoned arguments about the meaning of public opinion and the risks researchers are prepared to accept in the absence of a clear or stable ground truth.

Research on silicon respondents sits at a normative and methodological crossroads. Fully replacing human respondents in efforts to measure public opinion is difficult to imagine, but so, too, is ignoring the growing tension between the risks of relying exclusively on the shrinking subset of people who still answer surveys and the risks of incorporating silicon responses. This tension is likely to intensify as AI models continue to improve in their ability to generate human-like response profiles, even as their performance varies in opaque and sometimes unpredictable ways across models, model releases, prompting strategies, pre-training regimes, and post-training calibration.

The implications of this ambiguity extend well beyond technical considerations. Public opinion research has long been justified as a way of elevating human voices for normatively important projects related to representation, accountability, and social flourishing. Reliance on silicon respondents raises the possibility that silicon-based public opinion reflects echoes of past data or traces of online data, rather than the opinions of contemporary publics, in ways that privilege the most vocal and historically overrepresented opinions. If silicon respondents are used to measure the opinions of publics that are no longer easily or directly heard using conventional methods, the risk is not merely measurement error, but a shift in whose voices are counted that may amplify and homogenize dominant perspectives while further obscuring the views of the hardest-to-reach populations.

None of these concerns imply that silicon respondents should necessarily be rejected. But they do imply that their use is as much a theoretical claim as a technical choice – one requiring arguments about the conception of public opinion being assumed, what validation against past human survey responses can and cannot establish, and what epistemological and statistical risks researchers are willing to accept when no stable ground truth exists. As the technological capacity to generate silicon public opinion data increases, the central challenge is less about the technology and more about how its use affects the normative commitments that motivate the study of public opinion.

IV. Transparency, Disclosure and Replication

As previously noted, transparency and disclosure are central to the credibility of survey research (Jamieson et al., 2023). Even conventional survey replication is approximate rather than exact, given the inherent variability of human responses. As a result, survey research places a premium on transparency and the clear reporting of study design, recruitment, measurement, and analytic choices to enable independent scrutiny, meaningful replication, and appropriate interpretation. Yet even in conventional survey research, norms of disclosure can vary, and documentation is often not standardized.

The integration of AI further complicates traditional notions of replication and reproducibility (see the Barrie et al., chapter in this volume). As with survey research itself, replication of AI-integrated tools or data cannot mean simply repeating the same procedure and obtaining identical results because models are frequently retrained and altered without notice, and stochastic generation introduces variation across runs due to random seeds, parameter settings, or minor prompt changes beyond the researcher's control.

These characteristics raise a deeper conceptual challenge: What does replication mean when both the tool and its internal states evolve over time? Perfect replication is extremely difficult when studying digitally-mediated phenomena, due to the problem of temporal validity (Munger, 2023). As more politically relevant behavior, including the measurement of public opinion, incorporates generally non-deterministic AI throughout the research process, the problem of temporal validity worsens and the notion of replication shifts from a standard of duplication toward documenting

conditions of use and assessing robustness across plausible configurations. Taken together these issues shift the focus towards reproducibility of the process rather than replication of the exact results.

The problem is further intensified when research workflows depend on proprietary systems with opaque data curation and model update practices. As prior experience with corporate APIs (Freelon, 2018), industry collaborations (Munger, 2025), and online convenience samples (Ternovski and Orr, 2022) has shown, reliance on privately controlled infrastructures can render replication infeasible and undermine cumulative knowledge production. There is little reason to assume that widespread access to proprietary LLMs offered at heavily subsidized cost will persist over time. There is also the troubling possibility that model behavior may be shaped by commercial, political, or strategic incentives that are difficult for researchers to observe or diagnose or anticipate. While overt interventions may be easy to avoid, more subtle forms of influence are likely to be far harder to detect (e.g., Waight et al., (2026) shows that US frontier LLMs have already been influenced by Chinese state media control).

Although some degree of opacity may be unavoidable given the scale and complexity of the data now under study, concerns about replication and transparency must remain central to the study of public opinion. To that end, Spirling (2023) calls for the embrace of open-source LLMs, and Palmer et al. (2024) ask that researchers who incorporate proprietary LLMs explicitly justify their decision to do so. These proposals emphasize that transparency in AI-assisted research is not merely a technical concern, but an institutional and epistemic one, requiring deliberate choices about which tools are used and how their limitations are communicated. Cheng and colleagues (2023), motivated by concerns over how to evaluate the quality of silicon respondents generated using any type of LLM, proposed the CoMPosT framework suggesting disclosures of context, model, persona and topic of simulated respondents.

A commitment to transparency is perhaps especially important when using AI in the study of public opinion given the close connection between public opinion and core principles of democratic accountability and political representation. Measures of public sentiment are routinely used to evaluate whether governments are responsive to citizens, to justify policy choices, and to assess the legitimacy of political outcomes. Because public opinion data play such a central role in characterizing the relationship between societies and their governing institutions, and because AI can embed consequential assumptions at every stage of the research process, opacity about how those data are generated, processed, or interpreted carries particularly high stakes. Without adequate transparency, tools designed to describe public sentiment risk becoming mechanisms that redefine it, with consequences for how responsiveness, representation, and legitimacy are evaluated or conveyed.

Developing shared norms and standards for disclosing AI use in public opinion research, analogous to existing disclosure frameworks in survey research, is an important goal, although doing so is beyond the scope of this chapter (but see, for example, the CoMPosT proposal by Cheng et al. (2023), the *AAPOR Task Force on Responsible AI Integration in Survey Research Report* (Rothschild et al., 2026)**,** and the *AAPOR Code of Professional Ethics and Practices*). While a broader, ongoing conversation is essential, our focus is to elevate some important considerations for such a conversation (see also Barrie et al., 2025).

We recognize that the increasing integration of AI will make it difficult to document every use of AI across all survey research workflows. Nevertheless, it is essential that, at a minimum, researchers document how their use of AI could plausibly affect the data, the estimand, and the resulting conclusions. The goal is to provide sufficient transparency to allow others to trace, understand, and, when possible, replicate or at least partially reproduce the processes used to generate AI-assisted

responses or analyses so that others may evaluate how AI shapes knowledge claims being made. At the same time, researchers may be unaware of how AI systems shape outputs (Cheng et al., 2023), reinforcing the need for transparency standards that support independent evaluation.

In disclosing the use of AI, two distinct forms of transparency are especially important. The first is *technical transparency*, which concerns how AI systems were used in practice. At a minimum, this requires disclosing the date, model type, version, configuration parameters, and the exact prompts used to generate model outputs, as well as the nature of any pre-training or reinforcement learning applied. This form of transparency documents how the technical properties of specific models and implementations affect results, analyses, and conclusions.

The second form is *conceptual or interpretive transparency*, which concerns how researchers conceptualize and justify their use of AI. Beyond purely efficiency-enhancing applications, such as assistance with formatting, code writing, or minor editing, the use of AI is not a neutral technical detail. Instead, it constitutes a theoretical and methodological commitment that warrants justification proportional to its potential inferential consequences. As we have argued, when AI tools shape what data are collected, how responses are generated or interpreted, or which quantities are ultimately estimated, they embed assumptions about what public opinion is, how it can be observed, and what sources of error are acceptable. At a minimum, researchers should be explicit about which of the three targets their use of AI is aimed at: observed sample responses, latent respondent opinion, or population-level inference. This choice determines what fidelity means, what benchmarks are appropriate, and what claims the resulting data can legitimately support. These assumptions also condition the relationship between observed data and the substantive claims made about publics, attitudes, and behavior. Researchers should clearly describe where AI substitutes for or reshapes human input, what assumptions any substitution entails, and how it could plausibly affect substantive conclusions. In this sense, incorporating AI into public opinion research more closely resembles adopting a modeling framework than adopting a new software tool and, therefore, demands explicit and transparent justification of its conceptual foundations, as well as the statistical and epistemological risks it entails and its fitness for use.

V. Outstanding Questions

The discussions of this piece raise more questions than answers, but several questions seem particularly worthy of further attention by political scientists. In addition to a multitude of interesting technological and methodological questions associated with the implementation of AI in the study of public opinion, more enduring conceptual questions include:

- As AI increasingly shapes, mediates, or proxies for human beliefs, does this alter what we mean by public opinion, or merely how we observe it?
- How should we interpret LLM-based summaries of human expression? On the one hand they expand the information that can be used to characterize human opinion beyond survey responses, but on the other hand it is unclear how those expressions relate to human populations of interest.
- Given that LLMs overrepresent text-rich groups and high-resource languages, do AI-generated summaries or silicon respondents amplify the voices of the already-empowered? If so, how should researchers account for these distortions?

- As work continues to evaluate the performance of silicon respondents as a proxy for human opinion, what are the implications of the fact that available human survey benchmarks are themselves historical, unstable, and subject to survey error?
- What norms should govern the use and disclosure of AI in opinion measurement and political communication? How can political science maintain scientific norms of disclosure and replication when outputs result from opaque and changing processes?
- Normatively, what happens to democratic legitimacy when the line between measuring and manufacturing opinion collapses? Particularly as silicon respondents and AI-based persuasive systems are deployed at scale?
- Does the field need a unified framework that is analogous to the Total Survey Error paradigm to characterize the myriad ways that AI may affect the production, expression, and measurement of public opinion? AI introduces new forms of possibly correlated error at every stage of the opinion lifecycle – from opinion formation to generation and expression to measurement.

VI. Conclusions and Implications

In the novel *The Hitchhiker's Guide to the Galaxy*, the supercomputer Deep Thought famously reports that the answer to the Ultimate Question of Life, the Universe, and Everything is merely "42." This concise and precisely measured answer was provided with no explanation, and the task then became one of understanding what that answer meant. Solving that deeper problem required consulting not a machine, but rather building an even more complex system – Earth – as a planet-sized apparatus populated with human beings whose lived experiences, interactions, and observations were essential inputs to discovering the underlying question to which "42" was the answer.

Our contemporary engagement with AI and public opinion increasingly resembles this fictional dilemma. Large language models generate outputs that appear authoritative, nuanced, and complete, often with remarkable fluency and internal coherence. Yet the processes that produce those outputs grow more opaque as systems scale, training data expand, and alignment mechanisms multiply. As with Deep Thought, the danger is not that the answers are wrong in any simple sense, but that their apparent completeness invites interpretation without understanding. And just as Deep Thought required an entire planet of human experience to give meaning to "42," the outputs of AI systems demand human theory, judgment, and contextual grounding if they are to be meaningfully interpreted as claims about public opinion.

This tension is not merely technical. Throughout this chapter, we have argued that AI does not simply offer new tools for measuring public opinion; it unsettles the boundary between measurement and construction itself. When AI systems summarize human expression, impute missing responses, simulate respondents, or generate an entire silicon public, they implicitly encode assumptions about what public opinion is, whose voices matter, and which forms of expression count as evidence. These assumptions are often hidden in training data, prompts, model architectures, or proprietary updates, but they nonetheless shape the inferences researchers draw and the conclusions about public opinion that can circulate far beyond academia. The result is a growing risk that outputs optimized for plausibility, efficiency, or scalability may come to stand in for publics that are no longer directly heard.

The paradox is that as AI becomes easier and more capable, our responsibility as scholars only increases. Greater computational power does not absolve researchers from grappling with theory,

uncertainty, and normative judgment; it intensifies those obligations. Public opinion research has long been justified not only as a scientific enterprise, but as a democratic effort to make the preferences, beliefs, and concerns of citizens visible and consequential. If AI systems increasingly mediate, approximate, or replace those voices, the central question is not whether the technology works but whether its use preserves the epistemic and democratic commitments that have motivated the study of public opinion.

Seen in this light, the challenge posed by AI is not to extract ever more precise answers from increasingly powerful systems, but to remain attentive to the questions those answers are taken to resolve. Without sustained attention to theory, transparency, and human grounding, we risk mistaking fluent output for understanding and precision for legitimacy. As with Deep Thought, the most important work lies not in generating answers but in ensuring that we still know what we are asking, why we are asking it, and whose voices those answers are meant to represent.

*****

AI Use Disclosure: Generative AI tools were used for minor editorial assistance, including limited language editing to reduce redundancy and formatting references; all substantive content, analysis, and interpretations are the authors' own.

References:

Adams, D. (1980). *The Hitchhiker's Guide to the Galaxy*. New York: Harmony Books.

Alshaabi, T., Adams, J. L., Arnold, M. V., Minot, J. R., Dewhurst, D. R, Reagan, A. J., Danforth, C. M., and Dodds, P. S. (2021). Storywrangler: A Massive Exploratorium for Sociolinguistic, Cultural, Socioeconomic, and Political Timelines Using Twitter. *Science Advances,* 29(7). https://doi.org/10.1126/sciadv.abe6534

Amirova, A., Fteropoulli, T., Ahmed, N., Cowie, M. R., & Leibo, J. Z. (2024). Framework-Based Qualitative Analysis of Free Responses of Large Language Models: Algorithmic Fidelity. *PLOS ONE*, 19(3), 1–33. https://doi.org/10.1371/journal.pone.0300024

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., Evans, J., & Bernstein, M. S. (2025). Position: LLM Social Simulations Are a Promising Research Method. In *Forty-second International Conference on Machine Learning Position Paper Track*. https://proceedings.mlr.press/v267/anthis25a.html

Argyle, L. P. (2025). Political persuasion by artificial intelligence. *Science*, 390, 983–984. https://doi.org/10.1126/science.aec9293

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 337-351. https://doi.org/10.1017/pan.2023.2

Argyle, L. P., Busby, E. C., Gubler, J. R., Hepner, B., Lyman, A., & Wingate, D. (2025). Arti-'fickle' intelligence: using LLMs as a tool for inference in the political and social sciences. *Nature Computational Science*, 5, 737–744. https://doi.org/10.1038/s43588-025-00843-4

Arrow, K. J. (1951). *Social Choice and Individual Values*. New York: John Wiley & Sons.

Ashokkumar, A., Hewitt, L., Ghezae, I., & Willer, R. (2025). Predicting Results of Social Science Experiments Using Large Language Models. Online preprint. https://docsend.com/view/ity6yf2dansesucf

Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J., & Hemphill, L. (2024). Prompt Design Matters for Computational Social Science Tasks but in Unpredictable Ways. arXiv preprint. arXiv:2406.11980 https://doi.org/10.48550/arXiv.2406.11980

Baack, S. (2024). A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *Proceedings of The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2199–2208. https://doi.org/10.1145/3630106.3659033

Bailey, M. A. (2024)**.** *Polling at a Crossroads: Rethinking Modern Survey Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108697798

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Barrie, C., Palmer, A., & Spirling, A. (2025). Replication for Language Models. Problems, Principles, and Best Practices for Political Science. Working paper. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf.

Behrend, T. S., & Landers, R. N. (2025). Participant interactions with artificial intelligence: Using large language models to generate research materials for surveys and experiments. *Journal of Business and Psychology*, 40(6), 1275–1297. https://doi.org/10.1007/s10869-025-10035-6

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Berinsky, A. J. (2004). *Silent Voices: Public Opinion and Political Participation in America*. Princeton University Press.

Bisbee, J., Clinton, J.D., Larson, J. M & Lee, D. I. (2026). AI Pandering: Constructing Diverging Political Realities through Conversation. Working Paper.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, *32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Bisbee, J., & Spirling, A. (2025, March 14). *What To Do When Your LLM is Not State of the Art. When (not) to Worry About Misclassification and How to Correct for It in Social Science Applications* [Paper presentation]. 7th Annual Likert Symposium Generating and Classifying Text: Challenges and Benefits of Using Language Models in Social Research, University of Michigan, MI. https://surveydatascience.isr.umich.edu/wp-content/uploads/2025/01/James-Bisbee-SLIDES-Reduced-Likert-2025.pdf

Blumer, H. (1948). Public Opinion and Public Opinion Polling. *American Sociological Review*, 13(5), 542–549.

Bryce, J. (1888)**.** *The American Commonwealth*. Vol. II. New York: The Macmillan Company.

Burke, E. (1774). *Speech to the Electors of Bristol*. In *The Works of the Right Honourable Edmund Burke*, Vol. 2. London: F. & C. Rivington.

Buskirk, T. D., Keusch, F., von der Heyde, L., & Eck, A. (2025a). More Parameters Than Populations: A Systematic Literature Review of Large Language Models within Survey Research. arXiv preprint arXiv:2509.03391. https://doi.org/10.48550/arXiv.2509.03391

Buskirk, T.D., Eck, A., & Timbrook, J. (2025b). The Task Is to Improve the Ask: An Experiment for Developing Prompts to Generate High Quality Survey Items from Large Language Models. *http://dx.doi.org/10.2139/ssrn.5377878*

Buskirk, T.D., Eck, A., Timbrook, J., & Tatum, H. (2025c, May 14-16) *Is Your Chatbot Smarter Than a 5th Grader? An Experiment Testing the Steerability of Reading Levels of Survey Questions Created Using Generative AI Tools* [Paper presentation] 80th Annual American Association of Public Opinion Research Conference, Saint Louis, MO. https://aapor.confex.com/aapor/2025/meetingapp.cgi/Paper/3980

Caliskan, A., Bryson J. J., & Narayanan, A. (2017). Semantics Derived Automatically From Language Corpora Contain Human-like Biases. *Science*, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cerina, R., & Duch, R. (2025). The 2024 US Presidential Election PoSSUM Poll. *PS: Political Science & Politics*, *58*(2), 286–297. https://doi.org/10.1017/S1049096524000982

Cheng, M., Tiziano P., & Yang, D. (2023). CoMPosT: Characterizing and evaluating caricature in LLM simulations. arXiv preprint arXiv:2310.11501. https://arxiv.org/abs/2310.11501

Chintakunta, S. S., Nascimento, N., & Guimaraes, E. (2025). Large language models in the data science lifecycle: A systematic mapping study. arXiv preprint arXiv:2508.11698. https://arxiv.org/abs/2508.11698

Clinton, J. D., Agiesta, J., Brenan, M., Burge, C., Connelly, M., Edwards-Levy, A., Fraga, B., Guskin, E., Hillygus, D. S., Jackson, C., Jones, J., Keeter, S., Khanna, K., Lapinski, J., Saad, L., Shaw, D., Smith, A., Wilson, D., & Wlezien, C. (2021). Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls. American Association of Public Opinion Research. https://aapor.org/wp-content/uploads/2022/11/AAPOR-Task-Force-on-2020-Pre-Election-Polling_Report-FNL.pdf

Converse, P. E. (1964). The Nature of Belief Systems in Mass Publics. In Campbell, A., Converse, P. E., Miller, W. E., & Stokes D. E. (Eds.), *The American Voter* (pp. 206–261). New York: John Wiley & Sons.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714), eadq1814. https://doi.org/10.1126/science.adq1814

DellaPosta, D. (2020). Pluralistic Collapse: The "Oil Spill" Model of Mass Opinion Polarization. *American Sociological Review*, *85*(3), 507-536. https://doi.org/10.1177/0003122420922989

DellaVigna, S., Pope D., & Vivalt, E. (2019). Predict Science to Improve Science: Systematic Collection of Predictions of Research Findings Can Provide Many Benefits. *Science* 366(6464), 428–429. https://doi.org/10.1126/science.aaz1704

de Sola Pool, I., Abelson, R. P., & Popkin, S. L. (1964). *Candidates, Issues, and Strategies: A Computer Simulation of the 1960 Presidential Election.* (Rev. ed.). Cambridge: MIT Press.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI Language Models Replace Human Participants? *Trends in Cognitive Sciences*, 27(7), 597-600. https://doi.org/10.1016/j.tics.2023.04.008

Dillman, D. A. *Mail and Internet Surveys: The Tailored Design Method.* New York: Wiley, 2000.

Drechsler, J., & Haensch, A. (2024). 30 Years of Synthetic Data. *Statistical Science*, 39(2), 221-242. https://doi.org/10.1214/24-STS927

Duch, R., Jimenez, A., & Kotlarz, P. (2024). Improving Sampling and Generalizability in Field Experiments using Targeted Multi-Mode Convenience Samples and MRP. *Synthetic RCT, Talking to Machines*. https://talkingtomachines.org/projects/synthetic-rct/

Duch, R., Kotlarz, P., Low, R., Ohara, K., & Manning, B. S. (2024). *Draft: Artificially Intelligent RCT Pilot: Afro-Barometer and Candour II.* [Unpublished manuscript].

Fishkin, J. S. (1991). *Democracy and Deliberation: New Directions for Democratic Reform*. New Haven, CT: Yale University Press.

Freelon, D. Computational research in the post-API age. *Political Communication*, 35(4), 665-668. https://doi.org/10.1080/10584609.2018.1477506

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3), 1097-1179. https://doi.org/10.1162/coli_a_00524

Gallup, G. (1940) *The Pulse of Democracy: The Public-Opinion Poll and How It Works*. New York: Simon and Schuster.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2009) *Survey Methodology*, 2nd ed. Hoboken, NJ: Wiley.

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121, e2403116121. https://doi.org/10.1073/pnas.2403116121

Hackenburg, K., Tappin, B. M., Hewitt, L., Röttger, P., Hale, S., & Margetts, H. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6783), eaea3884. https://doi.org/10.1126/science.aea3884

Hause, L., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading Voters Using Human–Artificial Intelligence Dialogues. *Nature*, 648, 394-401. https://doi.org/10.1038/s41586-025-09771-9

Herbst, S. (1993). *Numbered Voices: How Opinion Polling Has Shaped American Politics*. Chicago: University of Chicago Press.

Hernandez, I., & Nie, W. (2023). The AI‐IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, *76*(4), 1011-1035. https://doi.org/10.1111/peps.12543

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, 633, 147–154. https://doi.org/10.1038/s41586-024-07856-5

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? [NBER Working Paper 31122]. https://doi.org/10.3386/w31122

Huang, Y., Wu, R., He, J., & Xiang, Y. (2024). Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R. *Journal of Global Health*, *14*, 04070. https://doi.org/10.7189/jogh.14.04070

Igo, S. E. (2007) *The Averaged American: Surveys, Citizens, and the Making of a Mass Public*. Cambridge, MA: Harvard University Press.

Iyengar, S., & Kinder, D. R. (1987). *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.

Jamieson, K. H., Lupia, A., Amaya, A., Brady, H. E., Bautista, R., Clinton, J. D., Dever, J. A., Dutwin, D., Goroff, D. L., Hillygus, D. S., Kennedy, C., Langer, G., Lapinski, J. S., Link, M., Philpot, T., Prewitt, K., Rivers, D., Vavreck, L., Wilson, D. C., & McNutt, M. C. (2023). Protecting the Integrity of Survey Research. *PNAS Nexus*, 2(3), pgad049. https://doi.org/10.1093/pnasnexus/pgad049

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2016). Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support. *Social Science Computer Review*, *35*(3), 336-356. https://doi.org/10.1177/0894439316631043

Kennedy, C., Blumenthal, M., Clement S.,, Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff L.,, Olson, K., Rivers, D., Saad, L., Witt, G. E., & Wlezien, C. (2018). An Evaluation of the 2016 Election Polls in the United States: AAPOR Task Force Report. *Public Opinion Quarterly*, 82(1), 1-33. https://doi.org/10.1093/poq/nfx047

Key, V.O. (1965). *Public Opinion and American Democracy*. Alfred A. Knopf: NY, NY.

Kim, J., & Lee, B. (2023). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. arXiv preprint. arXiv:2305.09620. https://arxiv.org/abs/2305.09620

Klašnja, M., Barberá, P., Beauchamp, N., Nagler, J., Tucker, J. A. (2017) Measuring Public Opinion with Social Media Data. In L. R. Atkeson, & R. M.l Alvarez (Eds.), *The Oxford Handbook of Polling and Survey Methods* (pp. 555-582). Oxford Handbooks. https://doi.org/10.1093/oxfordhb/9780190213299.013.3

Kozlov, M. (2025). AI chatbots can sway voters with remarkable ease — is it time to worry? *Nature*. https://doi.org/10.1038/d41586-025-03975-9

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Columbia University Press.

Lee, S., Peng, T., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., & Leiserowitz, A. (2024). Can Large Language Models Estimate Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias. *PLOS Climate*, 3(8), e0000429. https://doi.org/10.1371/journal.pclm.0000429

Lee, Sunghee, Tian, J. and Morales, S. (2025, November 21-22). *Evaluation of AI-Assisted Survey Questionnaire Translation* [Paper presentation]. 50th Annual Midwest Association of Public Opinion Research Conference, Chicago, IL.

Lepore, J. (2020). *If Then: How the Simulmatics Corporation Invented the Future*. New York: Liveright Publishing Corporation.

Li, C., & Qi, Y. (2025). Toward Accurate Psychological Simulations: Investigating LLMs' Responses to Personality and Cultural Variables. *Computers in Human Behavior*, 170, 108687. https://doi.org/10.1016/j.chb.2025.108687

Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, 648(8093), 394–401. https://doi.org/10.1038/s41586-025-09771-9

Lippert, S., Dreber, A., Johannesson, M., Tierney, W., Cyrus-Lai, W., Uhlmann, E. L., & Pfeiffer, T. (2024). Can Large Language Models Help Predict Results from a Complex Behavioural Science Study? *Royal Society Open Science*, 11(9), 240682. https://doi.org/10.1098/rsos.240682

Lippmann, W. (1922). *Public Opinion*. New York: Harcourt, Brace.

Lyman, A., Hepner, B., Argyle, L. P., Busby, E. C., Gubler, J. R., & Wingate, D. (2025). Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research. *Sociological Methods & Research*, 54(3), 1110-1155. https://doi.org/10.1177/00491241251342008

Madden, E. R. (2025). Evaluating the Use of Large Language Models as Synthetic Social Agents in Social Science Research. arXiv preprint. arXiv:2509.26080. https://arxiv.org/abs/2509.26080

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale. *Research & Politics*, 11(1). https://doi.org/10.1177/20531680241231468.

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media + Society*, 5(3). https://doi.org/10.1177/2056305119859294

Munger, K. (2020). All the news that's fit to click: The economics of clickbait media. *Political Communication*, 37(3), 376-397. https://doi.org/10.1080/10584609.2019.1687626

Munger, K. (2023). Temporal validity as meta-science. *Research & Politics*, 10(3). https://doi.org/10.1177/20531680231187271

Munger, K. (2024). *The YouTube Apparatus*. Cambridge: Cambridge University Press.

Munger, K. (2025). What Did We Learn About Political Communication from the Meta2020 Partnership? *Political communication*, 42(1), 201-207. https://doi.org/10.1080/10584609.2024.2446351

Olson, K., & Buskirk, T. D. (2025). "ChatBot" is a Two Syllable Word...Or Is It?: Using Generative AI for Survey Question Readability Assessments. *International Journal of Market Research*, 68(1), 61-81. https://doi.org/10.1177/14707853251389789

Padgett, Z., Maiorino, A. & Gutierrez, S. (2024, May 16). *Evaluating the Quality of Questionnaires Created with SurveyMonkey's Build with AI* [Paper presentation]. 79th Annual Conference of the American Association for Public Opinion Research, Atlanta, GA. https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/3198.

Page, B. I., & Shapiro, R. Y. (1992). *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. Chicago: University of Chicago Press.

Palmer, A., Smith, N. A. & Spirling, A. (2024). Using proprietary language models in academic research requires explicit justification. Nature Computational Science, 4(1), 2-3. https://doi.org/10.1038/s43588-023-00585-1

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative Agent Simulations of 1,000 People. arXiv preprint. arXiv:2411.10109. https://doi.org/10.48550/arXiv.2411.10109

Rothschild, D. M., Buskirk, T. D., Eckman, S., Hillygus, D. S., Kreuter, F., & Lazer, D. (2025). Successfully navigating the disruption AI will bring to survey research. *The Survey Statistician*, 92, 30–44.

Rothschild, D. M., Marlar, J., Amaya, A., Barari, S., Buskirk, T., Cobb, C., Gennai, J., Hillygus, D. S., Krupenkin, M., Lee, S., Steiger, D., Webb, B., & Korlakai Vinayak, R. (2026). *AAPOR Task Force on Responsible AI Integration in Survey Research Report*. American Association for Public Opinion Research.

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 15295–15323). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2024.acl-long.816

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). Towards coding social science datasets with language models. arXiv preprint. arXiv:2306.02177. https://doi.org/10.48550/arXiv.2306.02177

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, (pp. 29971-30004). https://dl.acm.org/doi/10.5555/3618408.3619652

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using Large Language Models to Generate Silicon Samples in Consumer and Marketing Research: Challenges, Opportunities, and Guidelines. *Psychology & Marketing*, 41(6), 1254–1270. https://doi.org/10.1002/mar.21982.

Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D. G., Ressa, M., Schweitzer, F., Song, D., Summerfield, C., Tang, A., Van Bavel, J. J., Van der Linden, S., & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy: The fusion of agentic AI and LLMs marks a new frontier in information warfare. *Science*, 391(6783), 354–357. https://doi.org/10.1126/science.adz1697

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Sepulvado, B. (2025, December). *Detecting AI Responses in Survey Data: NORC's Next Leap for Data Quality.* NORC at the University of Chicago. https://www.norc.org/research/library/detecting-ai-responses-survey-data-norcs-next-leap-data-quality.html

Settle, J. E. (2018). *Frenemies: How social media polarizes America*. Cambridge: Cambridge University Press.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Sniderman, P. M., Brody, R. A., & Tetlock, P. E. (1991). *Reasoning and Choice: Explorations in Political Psychology*. Cambridge: Cambridge University Press.

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, 616(7957), 413. *https://doi.org/10.1038/d41586-023-01295-4*

Tao, Y., Viberg, O., Baker, R. S., Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Ternovski, J., & Orr, L. (2022). A Note on Increases in Inattentive Online Survey-Takers Since 2020. *Journal of Quantitative Description: Digital Media*, 2. https://doi.org/10.51685/jqd.2022.002

Tiancheng, H., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2025). Generative Language Models Exhibit Social Identity Biases. *Nature Computational Science*, 5, 65–75. https://doi.org/10.1038/s43588-024-00741-1

Tirumala, S., Jain, N., Leybzon, D. D., & Buskirk, T. D. (2025, October 10). *Mic Drop or Data Flop? Evaluating the Fitness for Purpose of AI Voice Interviewers for Data Collection within Quantitative & Qualitative Research Contexts* [Paper presentation]. COLM 2025 NLPOR Workshop, Montreal, Canada. *https://openreview.net/pdf?id=Z4vRAcchxt*

Vargiu, C., & Nai, A. (2025). AI chatbots can persuade voters to change their minds. *Nature*, 648(8093), 287–288. https://doi.org/10.1038/d41586-025-03733-x

von der Heyde, L., Haensch, A. C., Weiß, B., & Daikeler, J. (2025). Using Large Language Models for Coding German Open-Ended Survey Responses on Survey Motivation. *Survey Research Methods*, 19(4), 355–370. https://doi.org/10.18148/srm/2025.v19i4.8568

Waight, H., Yang, E., Yuan, Y., Messing, S., Roberts, M., Stewart, B., & Tucker, J. (2026). State Media Control Influences Large Language Models. *Nature* [Forthcoming].

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3), 400–411. https://doi.org/10.1038/s42256-025-00986-z

Wang, Q., Wu, J.,, Jiang, Z., Tang, Z., Luo, B., Chen, N., Chen, W., He, B. (2025). LLM-based Human Simulations Have Not Yet Been Reliable. arXiv preprint. arXiv:2501.08579. https://doi.org/10.48550/arXiv.2501.08579

Westwood, S. J., Grimmer, J., & Hall, A. B. (2025). *Measuring Perceived Slant in Large Language Models Through User Evaluations* [Unpublished manuscript].

Wuttke, A., Aßenmacher, M., Klamm, C., Lang, M., Würschinger, Q. & Kreuter, F. (2025). AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 179-204). Albuquerque: Association for Computational Linguistics. https://aclanthology.org/2025.latechclfl-1.17.pdf

Xuechunzi, B., Wang, A., Sucholutsky, I., & Griffiths, T., L. (2025). Explicitly Unbiased Large Language Models Still Form Biased Associations. *Proceedings of the National Academy of Sciences of the United States of America*, 122(8). https://doi.org/10.1073/pnas.2416228122

Zaller, J. R. (1992)**.** *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.