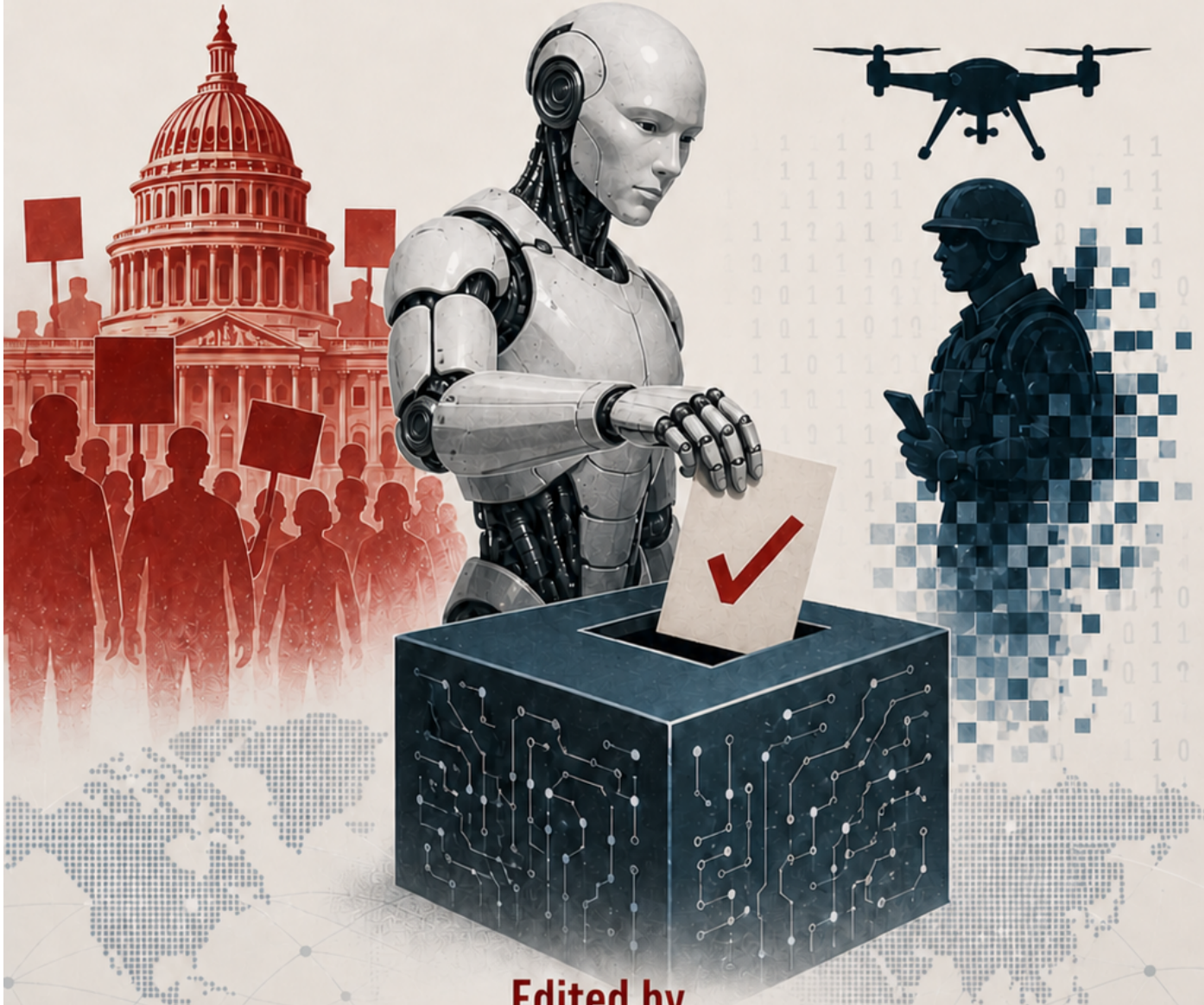
ARTIFICIAL INTELLIGENCE, POLITICS, AND POLITICAL SCIENCE



Edited by

NATHANIEL PERSILY

and JOSHUA A. TUCKER

Artificial Intelligence, Politics, and Political Science

Edited by Nathaniel Persily and Joshua A. Tucker

For Aaron, Noah, Sasha, and Mattie

Acknowledgements

Several people and institutions were instrumental in the production of this volume. Assembling sixty social scientists in the space of eight months to produce eleven chapters on an important, but fast-developing, topic is no easy feat. We would like to thank Taeku Lee for appointing us as co-chairs of the APSA Presidential Task Force on AI, as well as Jon Gurstelle and all the APSA staff who made this volume possible. We would like to thank Ho Ting "Adrian" Mak and Lisa Keen for editing and coordinating this volume, as well as our colleagues at Cambridge Press, including Jon Haslam and Carrie Parkinson, who shepherded this volume to publication. We would also like to thank the Stanford Law AI Initiative and NYU's Center for Social Media, AI, and Politics for providing institutional support.

Contents

[Note to typesetter: page numbers marked [x] to be inserted at typesetting stage.]

Figures

5

AI, Race, and Politics

Rachel Gillum, Gregory Leslie, and Cara Wong

Abstract: AI systems are reshaping racial and ethnic power dynamics across politics, governance, and scholarly inquiry, yet political science lacks systematic frameworks for analyzing when, where, and through what mechanisms these effects occur. This chapter surveys knowledge across three areas: 1) government use of AI in service delivery, surveillance, and coercive administration; 2) AI's impact on political information environments, mobilization, and electoral administration; and 3) AI as research infrastructure in the production of political knowledge. The methodology section examines how AI tools can introduce systematic distortions that standard disclosure practices do not address. In response, the chapter proposes an AI Measurement Statement (AIMS), a disclosure framework informed by practices in machine learning research and industry, designed to surface group-differentiated measurement risks and support transparency, construct validity, and cumulative knowledge production across political science subfields.

1. Introduction

Artificial intelligence is increasingly embedded in core political processes in ways that can reshape racial and ethnic patterns of representation and participation. Governments use algorithmic systems to allocate resources and enforce rules. Political campaigns deploy AI to shape information environments and mobilize voters. Scholars rely on AI tools to collect, code, and analyze political data. Across these domains, AI can reorder visibility, voice, and vulnerability, shaping who is recognizable to the state, who is targeted for outreach, how individuals navigate government services, and who can effectively challenge decisions made by opaque automated systems. These dynamics rarely operate in racially neutral ways, even when race is formally excluded from system design, because algorithmic systems draw on data, institutions, and political histories structured by racial and ethnic hierarchy.

The rapidly expanding use of AI across all facets of society raises foundational questions about power, participation, and accountability that cut across political science. However, existing research often treats AI either as a technical tool divorced from political context or as a generic normative problem of bias and ethics. This leaves underexamined how algorithmic systems interact with long-standing structures of racial and ethnic inequality, or how they reshape political science's own empirical foundations. Computer science identifies technical sources of group-differentiated performance and error, and advocacy organizations and journalists document discriminatory outcomes in particular settings. However, political scientists have produced comparatively little systematic research explaining when, where, and through what mechanisms AI systems reproduce, transform, or sometimes mitigate racial and ethnic political inequality.

This chapter has two aims in surveying available research on AI and racial and ethnic politics. First, it synthesizes insights across computer science, political science, and related fields regarding how algorithmic bias emerges and operates in political contexts. Second, it identifies gaps in current research on race and ethnic politics and proposes an agenda for studying how AI systems shape and interact with current theories of governance, political behavior, and scholarly inquiry. Much of the existing discussion on AI and race still relies on case studies, investigative reporting, and theoretical arguments about AI's potential benefits or harms, rather than research designs that establish causal mechanisms or specify the conditions under which particular effects occur. This gap poses risks for policy debates that proceed without evidence, but it also creates opportunities for scholars of race and ethnic politics to bring theories of power, identity, institutions, and mobilization to questions that are too often framed as purely technical.

Emerging scholarship suggests that bias in AI systems reflects structural features of data, institutions, and power relations, rather than isolated errors that can be corrected by technical fixes. Algorithmic systems, therefore, operate as both products of political inequality and mechanisms through which such inequality may be reproduced or transformed. Some work has documented how algorithmic systems reproduce or amplify racialized disadvantage, while other studies have identified contexts in which algorithmic decision making reduces certain forms of human discretion or increases consistency. Understanding which effects occur under what conditions, and how design choices, institutional context, and deployment practices shape outcomes, remains an open empirical question with implications for how political scientists understand representation, participation, state capacity, and accountability.

The chapter proceeds in three parts. First, it explains how racial bias becomes embedded in contemporary AI systems, clarifying why it persists across pretraining, alignment, and mitigation efforts even when race is formally excluded from model inputs. Second, it examines how AI systems operate across two linked domains – state governance and political participation. In government administration and coercive state functions, algorithmic tools can shape eligibility determinations, surveillance, and law enforcement in ways that may intensify racialized exclusion, even as some AI proponents emphasize potential gains in consistency or access under strong governance arrangements. In campaigns, political communication, and electoral administration, AI reshapes information environments, mobilization strategies, and participation burdens in ways that likely interact with organizational capacity and long-standing patterns of marginalization. Across these settings, scholars of race and ethnic politics are well positioned to analyze how algorithmic systems interact with group power dynamics, history, and institutional context.

Third, the chapter turns to political science methodology itself. As AI tools become integrated into research workflows – coding text, classifying observations, administering surveys, and summarizing literature – they function as measurement instruments whose error properties and representational biases shape inference and theory testing, often opaquely. Research in computer science and survey methodology shows that AI systems routinely infer sensitive attributes, exhibit systematic group-differentiated errors in measuring contested political constructs, and privilege majority perspectives in knowledge synthesis. These dynamics raise questions about construct validity, reproducibility, and epistemic authority. When AI systems are treated as neutral tools rather than as instruments requiring validation and disclosure, political scientists risk institutionalizing biased outputs as data and reshaping scholarly agendas under the appearance of neutrality. In response, the chapter proposes an AI Measurement Statement (AIMS), a disclosure standard adapted from model and system card practices in machine learning research and industry. It is designed specifically to surface group-differentiated measurement risks and support transparency, validity, and cumulative knowledge production across subfields, while recognizing the distinct analytical leverage that race and ethnic politics bring to evaluating algorithmic systems.

2. How Racial Bias Becomes Structural in AI Systems

Racial bias enters and persists throughout the development pipeline of large language models (LLMs), from pretraining to alignment to post-hoc mitigation. Despite extensive mitigation efforts, these models continue to reproduce racial inequality in systematic and predictable ways. Political scientists who study how technology reshapes racial hierarchy need to understand why bias emerges at scale and why it resists technical correction.

2.A Learning Inequality from the Internet

Racial bias enters the LLM pipeline at the pretraining stage, where models are trained on vast, internet-scraped corpora that disproportionately reflect white, Western, and higher-income linguistic norms that underrepresent or pathologize non-white speech, identities, and political perspectives (Bender et al. 2021; Blank 2017; Dodge et al. 2021; Noble 2018). Because these datasets are assembled from what is most readily available online – news media, books, social-media posts, forums, and code repositories – groups with greater digital visibility are overrepresented, while marginalized communities appear less frequently or in distorted contexts.

The consequences span modalities. Buolamwini and Gebru's *Gender Shades* study (2018) showed that facial analysis systems perform substantially worse on darker-skinned individuals due to nonrepresentative training data (Buolamwini and Gebru 2018; Kärkkäinen and Joo 2021; Raji and Buolamwini, 2019). Parallel dynamics appear in language models, with systems trained predominantly on English exhibiting higher error rates, weaker reasoning, and increased hallucination in other languages, limiting their reliability and safety for non-English speakers (Guo et al. 2025; Guerreiro et al. 2023; Qin et al. 2025).

Even in the absence of sampling bias, historical and institutional inequalities embedded in source texts shape model behavior. LLMs learn racialized associations not because they are explicitly instructed to do so but because such associations recur in news coverage, employment data,

policing records, and everyday online discourse. For example, models consistently associate Black-identifying names and African American English with negative attributes, mirroring long-standing patterns of discrimination in the underlying data (Blodgett and O'Connor, 2017; Blodgett et al., 2020; Caliskan et al., 2017). Audits of résumé-screening systems and job-ad delivery algorithms show that these learned associations can produce systematic disadvantages for applicants with Black-associated names, shaping both screening outcomes and access to employment opportunities (Wilson and Caliskan, 2024; Imana et al., 2021).

Efforts to remove biased content at the pretraining stage face structural limits. LLMs require broad, heterogeneous data to achieve generalizable performance, creating a tradeoff between scale and representational equity. Automated filters designed to remove toxic or abusive language disproportionately flag African American English because it co-occurs with racist abuse in training data (Davidson et al., 2019; Blodgett et al., 2020). Removing such content would reduce harassment but also Black political speech, cultural expression, and everyday language use, degrading model performance for Black users (Sap et al., 2019; Bender et al., 2021; Davidson et al., 2019; Hovy and Spruit, 2016). Eliminating explicit racial identifiers poses a similar tradeoff. Models rely on proxy features, such as dialect, vocabulary, geographic cues, educational institutions, and cultural references that correlate with race or gender and function as latent racial and gender markers (Bolukbasi et al., 2016; Sap et al., 2019). Stripping these proxy feature cues can degrade accuracy for marginalized groups, while preserving performance for dominant groups (Friedler et al., 2021). As a result, aggressive filtering often worsens representational disparities rather than resolving them. Racial bias in LLMs is therefore not a removable artifact but an emergent property of learning from racially stratified societies.

2.B Human Judgment as a Source of Bias

After pretraining, models undergo alignment through supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). Developers intend these stages to improve safety and usefulness, but each introduces additional pathways for racial bias.

In SFT, models are trained on curated examples of preferred responses. Multiple studies show that toxicity and safety datasets disproportionately label African American English and other marginalized dialects as offensive or unprofessional compared to semantically equivalent Standard American English (Blodgett and O'Connor, 2017; Davidson et al., 2019; Sap et al., 2019). This process teaches models to treat dominant linguistic norms as default and to suppress marginalized forms of expression.

Reinforcement learning from human feedback (RLHF) can compound these effects. In RLHF, human annotators – often contract workers operating under strict guidelines and time pressure – rank model outputs. The resulting reward models encode the cultural assumptions, linguistic preferences, and risk tolerances of both annotators and the platform designers who have hired them (Gray and Suri, 2019). Empirical work shows that RLHF can misrepresent minority viewpoints and induce "preference collapse," whereby optimization converges on dominant cultural norms while suppressing minority perspectives (Casper et al., 2023; Xiao et al., 2024). Thus, alignment processes intended to mitigate harm can inadvertently reinforce racial and cultural hierarchies.

2.C Why Technical Fixes Fall Short

Post-hoc debiasing techniques – such as output filtering, response constraints, or fine-grained parameter adjustments – address only surface model behavior. They cannot alter the deeper representations learned during pretraining and alignment, where racialized associations are embedded in the model's internal structure. As a result, disparities persist in downstream applications, even when overtly biased language is suppressed (Buolamwini and Gebru, 2018; Zhao et al., 2018).

These limitations are reinforced by mathematical constraints. Algorithmic fairness research has identified multiple distinct fairness criteria that appear desirable for prediction systems, including equal error rates across groups, calibrated predictions that reflect true risk levels, and demographic parity in outcomes. However, foundational work demonstrates that these criteria cannot be simultaneously satisfied except in trivial cases (Kleinberg et al., 2017; Pleiss et al., 2017). The incompatibility becomes particularly acute when groups exhibit different base rates of the outcome being predicted. For example, Chouldechova (2017) demonstrates that when Group A has a 10 percent recidivism rate while Group B has a 30 percent rate, equalizing false positive rates (incorrectly classifying low-risk individuals as high-risk) across groups necessarily produces unequal false negative rates (incorrectly classifying high-risk individuals as low-risk), and vice versa. Each fairness definition thus distributes prediction errors differently across demographic groups, creating unavoidable tradeoffs regarding which populations bear the costs of algorithmic mistakes.

Debates in computer science increasingly recognize that these are not merely technical problems, but sociotechnical ones. Treating fairness as an optimizable mathematical property obscures the fact that models learn from, and operate within, societies structured by racial inequality (Barocas et al., 2023; Selbst et al., 2019). From this perspective, bias mitigation cannot be reduced to parameter tuning. It requires attention to institutional choices, deployment contexts, and the political consequences of embedding algorithmic systems in unequal social worlds.

These technical dynamics have implications that extend well beyond model development. Because bias in large language models is structural rather than incidental, it shapes downstream effects across the political domains examined in this chapter – conditioning how algorithmic systems operate in government administration, immigration enforcement, and policing; how campaigns deploy AI for targeting and mobilization; and, how political information environments are curated and amplified. Scholars have characterized these dynamics as "techno-racism," referring to the ways in which technical systems encode and reproduce racial hierarchies while appearing neutral, objective, or efficient, thereby extending existing patterns of inequality through automated and institutionalized processes (Benjamin, 2019). Understanding how and why bias enters AI systems is therefore essential not only for evaluating governance and political consequences but also for assessing the reliability of the empirical evidence scholars use to study those consequences. The sections that follow examine these dynamics across institutional contexts and methodological approaches, identifying where AI reshapes political participation, representation, and state power, and where new research designs and disclosure standards are needed to ensure valid inference and cumulative knowledge production.

3. Linking AI to Political Outcomes

3.A AI, the State, and Racialized Governance

Governments now use AI and algorithmic systems in ways that touch the most consequential aspects of civic life (Overton, 2024). These systems determine who qualifies for public assistance, how agencies allocate scarce resources, how police and immigration authorities enforce the law, and how legislatures draw district lines to ensure equal representation or electoral advantage. Unlike most private-sector applications, government AI mediates access to rights, benefits, liberty, and political membership, and residents (citizens or not) often cannot exit or avoid these encounters. Policymakers (and critics) in the United States, the European Union, and elsewhere have responded by focusing heavily on public-sector AI, as algorithmic decision making shifts consequential judgments from accountable officials to technical systems that are difficult to interrogate, contest, or correct. Once embedded in state institutions, these systems reshape how governments classify, allocate, and coerce, often invisibly and with limited recourse for the people they classify. For scholars of race and ethnic politics, the critical task is to understand how these systems reorganize state capacity, discretion, and accountability, and whether they reproduce or reconfigure racialized governance (Omi and Winant, 1986).

3.A.1 Automated Administration and Racialized Burden

In public administration, AI and algorithmic systems are increasingly used to support eligibility determination, prioritization, and fraud detection in social services like welfare (Alon-Barkat, 2025; see also Chapter 4 of this volume, regarding public sector policy). Carefully designed and governed systems can reduce discretion, arbitrariness, and administrative error. They can lower transaction costs in ways that expand access for people who would otherwise be excluded from complex bureaucratic processes, though they can also create new barriers to accountability.

A growing body of evidence shows that algorithmic administration produces uneven and often racialized effects. Automated systems generate false denials, delays, and exclusion when they rely on rigid rules, incomplete administrative data, or error-prone matching. Virginia Eubanks's *Automating Inequality* (2018) synthesizes multiple case studies in which automated eligibility systems intensified hardship for poor and working-class populations by expanding surveillance, narrowing administrative discretion, and shifting the burden of proof onto applicants. Subsequent scholarship indicates that such harms frequently fall disproportionately on racial and ethnic minorities, both because these groups are overrepresented in means-tested programs and because historical inequalities are embedded in administrative data and decision criteria (Alon-Barkat, 2025; Kasy, 2024; Obermeyer et al., 2019). Research on algorithmic allocation and risk assessment further shows that optimization around organizational efficiency or fraud reduction can disadvantage marginalized groups even in the absence of explicit racial targeting (Benjamin, 2019; Timmons et al., 2022). Language access barriers compound these effects for many immigrant applicants navigating automated eligibility systems, particularly when machine translation introduces errors or when multilingual interfaces are unavailable for less-resourced languages (Hero, 1992).

Algorithmic systems now mediate access to housing at multiple stages, and the evidence of racial disparity is consistent across each. Machine learning mortgage models increase racial disparity in both approval rates and interest rates compared to traditional models (Fuster et al., 2022), and

nominally race-blind automated underwriting systems produce similar patterns (Bartlett et al., 2022). Algorithmic tenant screening compounds these effects in rental markets. What connects these outcomes is the data on which these systems draw: credit scores, eviction records, and criminal histories generated by the same discriminatory institutions that structured the inequalities these tools claim to measure neutrally, including redlining, racialized policing, and employment discrimination (Humber, 2023).

Administrative burden theory shows how procedural design structures learning, compliance, and psychological costs, with particularly strong effects for marginalized populations (Herd and Moynihan, 2018; Ray et al., 2022; Soss 1999). The central question is under what governance arrangements algorithmic administration expands access, and under what conditions does it reproduce or intensify racial and ethnic exclusion. Existing research documents the disparate impacts in specific policy domains, but systematic causal evidence across programs remains limited.

3.A.2 Risk, Surveillance, and Self-Reinforcing Classification

In coercive domains such as policing, criminal courts, and surveillance, AI could, in principle, reduce individual discretion and racial profiling. In practice, researchers and journalists have documented significant limitations. Computer science and criminology research demonstrates that predictive policing systems trained on historical enforcement data generate self-reinforcing feedback loops by intensifying patrols and arrests in already over-policed neighborhoods, regardless of underlying crime rates (Lum and Isaac, 2016; Ensign et al., 2018; Bennett Moses and Chan, 2018). Ethnographic and organizational research further demonstrates that data-driven policing does not eliminate discretion but redistributes it, concentrating judgment in earlier stages of classification and targeting, while rendering those judgments less visible to oversight (Brayne, 2017).

Risk assessment tools used in pretrial detention and sentencing raise related concerns. Investigative reporting and subsequent empirical analysis have documented racially disparate error rates in widely used tools (Angwin et al., 2016; Kleinberg et al., 2018; Skeem and Lowenkamp, 2016; Mayson, 2019). The politically salient issue is not which fairness metric is normatively correct, but how particular definitions become institutionalized through law, procurement, and bureaucratic practice. For race and ethnic politics, this shifts attention to how algorithmic classifications reshape racial meaning, criminalization, and perceived political belonging, even when race is formally excluded from model inputs (Epp et al., 2014; Lerman and Weaver, 2014).

Immigration and border control represent another high-stakes and comparatively opaque domain of government AI. Journalists and advocacy researchers have documented algorithmic risk scoring, biometric identification, and automated triage in visa processing, asylum screening, and border surveillance. They operate in institutional contexts marked by power asymmetries, overlapping jurisdictions, limited procedural protections, and restricted public visibility (Varsanyi, 2008; Varsanyi et al., 2011; Lee, 2019; McNamara and Tikka, 2023). Existing analyses argue that these technologies risk reproducing racialized exclusion by encoding assumptions about risk, credibility, and deservingness into automated decision processes. Emerging scholarship documents how data-driven border systems can entrench techno-racism and differential mobility control, even as systematic causal evidence remains limited (Molnar, 2019, 2024; Rinaldi and Teo,

2025). These systems disproportionately affect Latino and Asian American communities, and proximity to immigration enforcement creates documented spillover effects – reduced engagement with public programs, healthcare, and civic institutions – even among citizens and permanent residents not directly targeted (Pedraza et al., 2017; Ramakrishnan 2006).

3.A.3 Opacity, Contestation, and Power Asymmetries

A common thread runs through these domains: opacity and the structural difficulty of contestation. Legal scholars have shown that algorithmic systems can evade traditional civil rights scrutiny through proprietary protections, technical complexity, and fragmented responsibility across public and private actors (Barocas and Selbst, 2016; Kroll et al., 2017). For racially marginalized communities, these features compound existing vulnerabilities by limiting access to explanation and meaningful avenues for appeal, whether in eligibility determinations, risk scoring, or immigration adjudication. Research in political behavior demonstrates that perceptions of procedural unfairness reduce trust in institutions and willingness to comply with state authority (Tyler, 1990; Tyler et al., 1989; Wu et al., 2022; Kruis et al., 2023; Johnson et al., 2017), though direct causal evidence linking algorithmic systems specifically to group-differentiated trust outcomes remains limited.

This defines the central research agenda across all three domains. When do algorithmic systems widen power asymmetries between the state and citizens by centralizing expertise and obscuring decision logic? And when can transparency requirements, audits, or civil society interventions mitigate these effects (Mettler, 2011)? Under what institutional conditions can affected communities mobilize data, counter-models, or legal claims to challenge racially disparate outcomes? And when do legal, technical, or organizational barriers foreclose such challenges (Barocas and Selbst, 2016; Huq, 2019; Kim, 2022; Meng and DiSalvo, 2018)? How do migrants, advocates, and legal institutions challenge algorithmic decisions when evidence is inaccessible and appeal pathways are constrained? And when do litigation, audits, or investigative journalism succeed in reshaping policy? Addressing these questions requires integrating insights from public administration, political behavior, and racial politics to analyze AI – not as a neutral instrument of governance, but – as a reconfiguration of how the state classifies populations, exercises authority, and is held accountable.

3.B AI, Political Information, and Collective Action

Beyond direct state action, AI reshapes political power through information environments, mobilization, and electoral administration. Political science scholarship on media, participation in elections, and race suggests that these shifts from manual to automated AI processes will interact with existing racial inequalities rather than operate as neutral technological changes. Yet, the causal mechanisms and downstream effects remain underspecified (Besco, 2024; Jun et al., 2022; Flores and Coppock, 2018). This gap presents a significant research opportunity. The following section lays out three interrelated domains ripe for empirical examination around how AI restructures political information flows and exposure (See also Chapter 1 on Democracy and Chapter 2 in this volume on the Information Ecosystem).

3.B.1 Fragmented Information Environments

Political consultants and organizers increasingly use AI tools to generate, translate, and personalize political information. These tools offer genuine opportunities for expanding political access. AI-

assisted translation and plain-language summarization can reach voters in their native languages at a scale and cost previously unavailable to most campaigns and civic organizations, with particular potential for communities historically underserved by English-only political outreach. Automated tools can also help under-resourced campaigns and advocacy organizations communicate across linguistic communities that would otherwise require prohibitively costly translation infrastructure.

At the same time, the same capabilities enable campaigns and external actors to produce highly tailored misinformation or demobilizing content (Mauk and Grömping, 2024). Historically, such practices have disproportionately targeted minority communities to suppress turnout, sow confusion, and lower trust in election integrity (Uribe et al., 2025).

Accuracy and nuance present further complications. AI translation and summarization systems often perform worse for less prevalent languages and dialects, potentially altering message credibility or meaning (Fleisig et al., 2024). This affects not only Spanish-speaking Latino communities, but also Asian American populations whose native languages – Vietnamese, Tagalog, Korean, Chinese dialects, and more – may be poorly represented in training data. The result is that the communities with the most to gain from AI-assisted language access may also bear the highest risk of receiving mistranslated or culturally flattened political content. Similarly, AI-powered content moderation systems shape what political information remains visible to different communities, yet the conditions under which moderation decisions affect minority political speech differently from majority speech remain poorly understood (Oh and Downey, 2025).

Classical theories of electoral accountability assume that voters evaluate candidates and election officials within relatively shared information environments that allow coordination, comparison, and sanctioning (Fiorina, 1981; Arias et al., 2019). Algorithmic personalization disrupts this assumption. Different racial groups may observe fundamentally different information about what representatives said, did, or promised, especially if politicians explicitly tailor their messages depending on their audience (Glaser, 1996). A central research question for race and ethnic politics is whether algorithmically curated information environments strengthen racial group consciousness by reinforcing shared experiences, or whether they fragment political understanding in ways that impede collective accountability (Sanchez, 2006; Chong and Rogers, 2005). When minority communities receive systematically different political information through platform recommendations or campaign targeting, does this enhance within-group solidarity or instead prevent the cross-racial coalition formation that minority political influence often requires (Kaufmann, 2003)?

3.B.2 Targeting, Authenticity, and Organizational Capacity

Campaigns increasingly rely on AI-driven propensity modeling and microtargeting to allocate voter contact and outreach resources (Endres and Kelly, 2018; Savaget et al., 2019). A central question is whether these systems generate systematic differences in contact rates across racial and ethnic groups. Existing theory suggests such disparities emerge through optimization over historically skewed voter files rather than explicit campaign intent (Dong et al., 2025; Hersh 2015; Ross and Spencer, 2022). When AI systems are trained on voter files shaped by historical exclusion and suppression, they systematically classify voters in neighborhoods with lower recorded turnout

as low-return targets despite high latent potential for mobilization. Unlike traditional targeting, where field directors might recognize that low turnout reflects barriers rather than disinterest, algorithms optimize purely on statistical patterns in biased historical data, producing feedback loops that reinforce predictions of low propensity in future cycles (Barocas and Selbst, 2016). These same tools can also be repurposed explicitly for minority voter suppression (Panditharatne, 2024).

Algorithmic segmentation may also reshape collective political identities. The ability to detect intragroup heterogeneity rapidly – such as class or religious variation within racial groups – could enable campaigns to tailor appeals that weaken shared political identities and erode Linked Fate (Dawson, 1994; Cohen, 1999). While Linked Fate has been most extensively theorized for Black Americans, scholars have documented analogous processes of group consciousness among Latinos and Asian Americans, though with distinct mechanisms tied to immigration status, pan-ethnic identity formation, and language (Barreto and Segura, 2014; Junn and Masuoka, 2008; Gay et al., 2016). Platform algorithms that deliver racially segmented information environments could reduce opportunities for cross-racial exposure and shared framing, constraining multiracial coalition-building (Benjamin 2017). On the other hand, algorithmic analysis could help identify cross-cutting interests, such as common economic or educational concerns, that facilitate new coalition formation (Wong, 2006; Han, 2014). Whether AI-driven targeting fragments or reconfigures collective political identities depends on conditions that existing research has not yet specified.

Authenticity presents a distinct challenge for AI-mediated outreach. Field experiments consistently show that political contact increases turnout, with stronger effects when messengers share ethnic or cultural characteristics with voters (García Bedolla and Michelson 2012; Sinclair et al., 2013; Green and Gerber, 2019). AI can expand the scale and linguistic reach of such contact, as politicians can now recreate their voice and likeness across multiple languages, enabling outreach that would otherwise require prohibitive resources (Coltin, 2023). However, when AI-generated messages contain grammatical errors, culturally inappropriate phrasing, or stylistically flattened language, they may signal inauthenticity to recipients in ways campaigns cannot easily observe or correct. This creates a tradeoff between message volume and message quality whose net effects on minority political participation remain empirically unresolved.

These dynamics place community organizations in a pivotal role. When AI-generated outreach is perceived as inauthentic, trusted organizations serve as validators of political information and conduits for mobilization. At the same time, algorithmic optimization may reduce investment in these intermediaries by shifting resources toward individualized voter contact and away from neighborhood-based organizing (Hersh 2015; Kalla and Broockman, 2018). This suggests a potential feedback loop in which technical bias weakens direct outreach, while campaign strategy erodes the organizational infrastructure capable of compensating for those weaknesses. Platform moderation systems compound these pressures by flagging collective action messaging as spam or incitement at higher rates for racial justice organizing, increasing the cognitive and organizational costs of mobilization and pushing communities toward code-switching strategies to evade detection (Sap et al., 2019; Haimson et al., 2021; Lee et al., 2024).

3.B.3 Electoral Administration and Redistricting

Beyond affecting campaigns and information flows, AI increasingly mediates electoral administration itself. Election officials now rely on automation for signature matching, voter roll

purges, registration verification, and polling place allocation (Cable et al., 2023). When designed and governed carefully, AI-enabled systems can reduce administrative error, expand access, and standardize procedures in ways that benefit all voters, including racial and ethnic minorities. At the same time, these technologies introduce new participation risks that warrant close scrutiny.

Historical patterns of racial exclusion through administrative mechanisms such as literacy tests, poll taxes, and restrictive registration rules demonstrate that seemingly neutral procedures can nonetheless produce racially uneven effects (Behrens et al., 2003; Gray and Jenkins, 2025; Shah and Smith, 2021). Signature-matching algorithms exhibit variation linked to name structure and signature style, raising the possibility of disproportionate ballot rejection for minority voters (Blumenstein 2021). Automated voter roll maintenance systems may similarly flag minority voters as inactive at higher rates due to residential mobility associated with economic precarity and housing instability (Huber et al., 2021). While these disparities exist without AI, the scale, speed, and opacity of automated systems raise the stakes of differential error considerably.

Efficiency-oriented applications carry analogous risks. Algorithms used to allocate voting machines, poll workers, or early voting locations may reproduce historical inequalities if based solely on outdated turnout data. This can direct resources away from communities that were previously underrepresented, rather than correcting for that underinvestment (Stewart, 2013). AI-driven voter verification and profiling raise additional surveillance concerns. Minority and immigrant communities with histories of state scrutiny may reduce their willingness to register, vote, or organize, even in the absence of formal coercion (Pedraza et al., 2017; Farzan, 2018).

Taken together, AI-enabled election administration may reduce some participation burdens while introducing new ones. When errors are opaque and correction procedures complex, the burden of contestation shifts from the state to the individual. Communities shaped by voter suppression, surveillance, and bureaucratic exclusion face higher psychological and informational barriers to challenging these systems, allowing ostensibly neutral procedures to reproduce racial disparities in participation.

AI-assisted redistricting is among the most consequential applications of algorithmic tools in electoral politics because it enables both more precise gerrymandering and more rigorous detection of it. Computational algorithms now generate and evaluate millions of possible district maps, providing statistical baselines against which enacted maps can be assessed for partisan or racial bias (McCartan and Imai, 2023). These tools play a direct role in Voting Rights Act litigation, but Cho and Cain (2020) warn that their misuse poses a greater democratic threat than overt partisan map-drawing, because they allow gerrymandered outcomes to be validated as products of neutral computation. Computer-generated maps that optimize only for compactness and population equality can systematically underrepresent communities of color, because residential segregation – itself a product of racially discriminatory housing policy – means race-blind formal criteria do not produce race-neutral substantive outcomes (Chen and Rodden, 2013). The choice of which fairness criterion to encode (e.g. compactness, competitiveness, proportional representation, or VRA Section 2 compliance) is a political decision, and AI tools amplify the consequences of that choice across millions of voters simultaneously.

3.B.4 Feedback Loops and Unequal Participation

The dynamics examined in this section operate as interdependent forces, not isolated effects. Fragmented information environments weaken coordination and collective capacity. Diminished organizational strength heightens exposure to administrative barriers. And increased participation burdens erode the political power needed to contest or reshape the algorithmic systems producing those costs. These feedback loops are visible only through institutional analysis attentive to race, historical patterns of exclusion, and organizational capacity (Mettler and Soss, 2004; Pierson, 1993).

For scholars of race and ethnic politics, the core question is not whether AI affects minority political participation, but under what institutional, organizational, and regulatory conditions AI systems reproduce, reconfigure, or interrupt racialized political hierarchies. Addressing this question requires research designs that exploit institutional variation, examine staggered adoption of algorithmic systems, and document how communities resist, adapt to, or repurpose AI-mediated political environments (DaViera et al., 2024; Piccardi et al., 2025; Overton, 2026).

These dynamics reshape the *conditions* under which participation occurs, but they do not determine whether AI-related disruptions become objects of collective action or electoral accountability. Material harm does not automatically generate political salience (Chong, 1991). Whether AI-driven changes in surveillance, administrative burden, or information access translate into mobilization depends on attribution, framing, and politicization under conditions of inequality (Kinder and Sanders, 1996; Tesler., 2016; Hutchings and Valentino, 2004; Arnett, 2020). Specifying when and how those processes unfold is where theories of racial politics, administrative burden, and collective action intersect and where the subfield has the most to contribute.

4. AI as a Methodological Actor in Political Science

Alongside major consequences for real-world political behavior, AI systems are reshaping political science research itself. AI is increasingly embedded in research workflows, from data collection and survey design to content analysis and literature synthesis. In these roles, AI systems operate as methodological actors, systematically influencing – rather than neutrally reflecting – what scholars observe, how concepts are operationalized, and which patterns appear most salient.

The bias mechanisms identified in Section 2 manifest with particular force when AI is deployed as research infrastructure. Three structural features explain why. First, AI error is systematic rather than random; patterned by race, language, and dialect in ways that produce correlated measurement bias within groups, rather than noise that averages out. Second, these systems are opaque. Proprietary models update without notice, training data compositions remain undisclosed, and researchers often cannot determine whether observed patterns reflect substantive political differences or classification artifacts. Third, AI-generated outputs propagate through research pipelines, embedding early-stage errors into downstream findings in ways that can institutionalize racialized measurement artifacts as empirical fact. These concerns are not confined to race and ethnic politics. The companion methods chapter in this volume documents the same problem from a measurement standpoint, noting that LLM annotation errors may not be uniform throughout the data and that differential error across population subgroups can cause bias in downstream analysis that uses AI outputs as inputs (See Chapter 10, Barrie, Mellon, et al., this volume) While that

chapter addresses these risks as a general methodological problem, this chapter examines how they operate specifically across the racial, ethnic, and linguistic groups central to race and ethnic politics research.

The subsections that follow examine how these features manifest across specific research domains, from implicit racial inference and measurement validity to data collection, sampling, and scholarly knowledge production. Understanding these dynamics is essential not only for studying AI's impact on politics, but for ensuring that political science research does not inadvertently institutionalize the very disparities it seeks to explain.

4.A Implicit Racial Inference by AI Systems

A growing body of computer science research demonstrates that machine learning systems routinely infer sensitive attributes such as race and ethnicity implicitly, even when researchers do not ask them to and those attributes are not explicitly provided as inputs. Systems make these inferences through signals such as names, dialect, syntax, geography, topics, and network structure – features that function as racial proxies even when race is formally absent from the model (Kosinski et al., 2013; Caliskan et al., 2017; Elazar and Goldberg, 2018; Zhang, Lemoine, and Mitchell, 2018). The concern is not only that such inference occurs at scale, but that we often cannot observe how latent identity predictions shape downstream classifications and decisions (Barocas and Selbst, 2016).

When political scientists use AI systems to classify text, code events, predict behavior, or construct measures of political identity and opinion, those systems may be making implicit racial inferences that shape outputs in ways the researcher cannot observe or audit. A classifier that infers race from name, dialect, or geographic proxy and then treats racialized individuals differently in its outputs will introduce measurement error that is correlated with the very group characteristics the researcher is trying to study, producing biased estimates precisely where valid inference is most needed. Unlike random measurement error, which attenuates relationships and can be addressed partially through standard techniques, this form of error is directional and systematic, determined by racial assumptions embedded in training data, rather than by the theoretical relationships under investigation. The core methodological challenge is how researchers assess construct validity when the measurement instrument may be operationalizing race in ways they did not intend and cannot directly observe. This is a question that requires new tools and disclosure standards to address (Omi and Winant 1986, Burrell, 2016).

4.B Measurement Validity and the Automation of Political Constructs

A second methodological frontier concerns measurement validity when political scientists rely on pretrained classifiers to operationalize political concepts like extremism, toxicity, hate, misinformation, ideology, affect, trust, or grievance. The use of such models implicitly accepts the normative assumptions embedded in training data and labeling practices, which overwhelmingly reflect the cultural norms and priorities of majority groups (Noble, 2018; Benjamin, 2019).

For political science, the methodological risk is that highly contested political constructs may come to appear fixed and objective over time. When classifiers learn what counts as extremism, civility, or sentiment from historical data, they embed particular theories of politics into measurement

itself, often without explicit theoretical justification (Kiritchenko and Mohammad, 2018; Jacobs and Wallach, 2021). Used uncritically, such tools can reproduce racial asymmetries in measurement error that then propagate into substantive findings. The problem is compounded by opacity: Proprietary models update without notice, training data compositions remain undisclosed, and researchers studying minority political behavior often cannot determine whether observed patterns reflect substantive political differences or classification artifacts. These questions sit at the intersection of construct validity, racial power, and epistemology, and they remain underexplored in mainstream political methodology.

4.C Translation and Multilingual Research

The measurement validity problems described above extend to multilingual research contexts. Studies of Latino political behavior, immigrant incorporation, cross-national comparative work, and research on non-English-speaking minority communities all depend on the ability to analyze political content produced in languages other than English. As AI translation and multilingual Natural Language Processing (NLP) tools become standard research infrastructure, they introduce a category of differential error that the discipline has not yet adequately addressed.

As discussed earlier in this chapter, AI language systems perform substantially worse on languages other than English, and the performance gap is largest for the languages spoken by politically marginalized communities. This is a self-reinforcing mechanism whereby "high-resource" languages attract more training data, producing better models and more investment, while "lower-resource" languages fall further behind (Hovy and Prabhumoye, 2021). Sentiment analysis tools trained predominantly on English-language data risk misassigning emotional valence to Spanish and other minority-linked languages, and toxicity detection systems risk misclassifying culturally specific expressions of political intensity as harmful content (Sap et al., 2019; Blodgett et al., 2020). When a researcher uses AI to translate Spanish-language voter testimonials and then applies an English-trained sentiment classifier to the output, two distinct layers of differential error have been introduced before any analysis begins, neither visible in the final dataset.

The problem is not only technical accuracy. Language carries political meaning that systems trained on dominant-language corpora are poorly positioned to recover. Political speech in minority communities is often deliberate in its register, its use of in-group terminology, and its deployment of cultural reference (Rosa and Flores, 2017). Just as the discipline applies meaningful scrutiny to human translators – documenting credentials, assessing intercoder reliability, acknowledging interpretive choices – machine translation requires similarly careful examination.

4.D Estimating the Causal Effects of Race

AI systems also pose implications for how scholars estimate and understand the causal effects of race. Formally identifying treatment effects for race remains a particular challenge. Race is generally treated as immutable and assigned at birth (though see Agadjanian, 2022; Penner and Saperstein, 2008), making it difficult to credibly manipulate in experimental settings. Covariates commonly used to balance treatment assignment – education, income, neighborhood context – are themselves consequences of race, increasing the risk of post-treatment bias (VanderWeele and Hernán, 2012). Race is also a multidimensional construct encompassing a wide range of factors, for example, skin tone, eye and nose shape, dialect, socioeconomic status, and power relations (Sen and Wasow, 2016). This "bundle-of-sticks" character complicates causal inference by making it difficult to isolate which dimensions of race drive observed effects on social and political outcomes. Critically, many outcomes are driven by perceptions of race whose effects resist

decomposition into component parts, which limits how much any inferential strategy can recover (Baldus et al., 1983, Quillian et al., 2017, Harris and Findley, 2014).

The same inferential capacity that creates measurement risks in observational research also opens methodological possibilities when deployed deliberately and transparently in causal identification. Because machine learning models can infer race from names, dialect, geography, and social networks, they can decompose race into probabilistic signals, allowing researchers to estimate the distinct effects of its component features (Barocas et al., 2023; Imai and Khanna, 2016). Advances in AI-based causal modeling – double machine learning, random forests, gradient boosting – have also improved researchers' ability to adjust for the complex socioeconomic and spatial structures that shape racial inequality in observational settings (Dorie et al., 2019; Brand et al., 2023; Chernozhukov et al., 2018). That said, the impact of race exceeds the sum of its component parts, and these decomposition strategies are most credible for the subset of outcomes where perceptual and relational dimensions of race are not the primary drivers.

The more immediate risk is that algorithmic inference contaminates the identification strategies researchers use. AI's tendency to embed latent racial classifications may corrupt treatments, outcomes, and control variables simultaneously (Benthall and Haynes, 2019; Dressel and Farid, 2018), causing traditional identification strategies to conflate substantive causal effects with measurement artifacts generated by opaque computational systems (Cranmer, 2019; Grimmer et al., 2021). The result is not merely biased estimates, but a deeper erosion of the boundary between causal mechanisms and data infrastructure. Where the contamination problem can be identified, corrective approaches are emerging. Egami et al. (2023, 2024) develop design-based methods for downstream inference that adjust for imperfect AI surrogates, allowing researchers to account for non-uniform annotation error when AI outputs enter causal models. These methods address contamination only when its sources have already been identified, which returns the burden of proof to the documentation and validation practices taken up in the following sections, particularly in domains where identity is central to theory and inference.

4.E AI-Mediated Data Collection and Political Expression

The preceding subsections address bias in AI as a measurement instrument, discussing what happens to data once it has been collected. A distinct set of problems arises earlier in the research pipeline at the stage of data generation itself. AI systems increasingly mediate data collection by assisting with survey design, administering conversational surveys, probing responses, and summarizing qualitative material (Grimmer et al., 2021; Jurka et al., 2013; Cranmer, 2019; Stout and Garcia, 2022). Research in human-computer interaction and survey methodology suggests that conversational agents can affect disclosure patterns and social desirability bias, depending on perceived embodiment, anonymity, and institutional framing (Papneja and Yadav, 2024, Schuetzler et al., 2018, Ho et al., 2018, Xiao et al., 2020). Recent political science-adjacent work evaluates large language models as adaptive interviewers, showing both promise and variability relative to human interviewers (Wuttke et al., 2025).

These studies demonstrate that AI interviewers are not neutral actors. They impose a particular linguistic style, constrain permissible forms of expression, and deliver feedback signals that systematically shape participants' responses in ways that can further reinforce dominant norms (Ho et al. 2018). When administered by governments, universities, or platforms with documented

histories of monitoring marginalized communities, AI-powered data collection systems may carry surveillance associations that can alter respondent behavior before a single question is answered (Penney, 2016, 2022). For instance, Black Americans subjected to carceral surveillance show well-documented patterns of institutional avoidance and civic disengagement, including reduced willingness to interact with state-affiliated data collection (Lerman and Weaver, 2010, 2014). Latino communities in perceived anti-immigrant climates show reduced institutional engagement broadly, including those who are U.S.-born citizens who face no direct legal risk (Vargas et al., 2017; Asad, 2020). When AI mediates the survey environment, it may activate these existing threat perceptions, producing differential non-response that is systematic and racially correlated rather than random. AI-administered surveys thus risk compounding the very underrepresentation they purport to study.

For studies involving diverse populations, this introduces underexamined measurement concerns. AI-mediated interviewing may reshape political expression differently across racial groups, altering what respondents choose to disclose, how they frame experiences, or whether they self-censor. Does AI reduce traditional interviewer effects by standardizing interaction, or intensify them by introducing new forms of perceived monitoring? And through what mechanisms – language choice, persona design, institutional association – do these effects operate across racial groups? Addressing these questions is essential for understanding whose political voices are captured and how racialized experience enters the research record.

4.F AI-Mediated Sampling and Population Bias

The preceding subsection examined how AI shapes what respondents express once they are inside the data collection process; the problem considered here begins earlier, at the stage where researchers decide which populations are observable at all.

When political scientists build corpora by scraping social media platforms or querying platform APIs, the populations captured may not be representative, and the gaps can be racialized in systematic ways. Social media platforms carry substantial built-in population biases that researchers routinely fail to acknowledge or correct, and when AI tools mediate access, proprietary platform algorithms introduce additional filtering whose effects cannot be observed or audited by outside researchers (Ruths and Pfeffer, 2014; Olteanu et al., 2019).

Researchers who build corpora of political speech from social media platforms are sampling from user populations that differ systematically by race in their platform participation, posting behavior, and visibility (Barberá and Rivero, 2015; Auxier, 2020; Ruths and Pfeffer, 2014). AI-based query tools built on majority-language assumptions may not reliably retrieve political content expressed through African American Vernacular English, code-switching, or community-specific hashtag practices. This produces systematic exclusion – not missing data in the conventional sense, but communities rendered invisible by the design assumptions embedded in the research infrastructure used to observe them. Freelon, McIlwain, and Clark (2016) document the scale of what is at stake, showing that Black Twitter constituted a distinctive and consequential site of political organizing and agenda setting, precisely the kind of political activity that AI-mediated corpus construction is liable to undercount. As platform API restrictions have tightened since 2023, this problem has grown more acute and will require explicit methodological attention going forward.

4.G Algorithmic Distortion of Benchmarking Infrastructure

The preceding subsections document how AI systems introduce measurement distortions at multiple stages of the research pipeline. A less visible problem concerns the benchmarking data researchers use to detect and correct those distortions. Scholars routinely rely on Census-derived population estimates to evaluate sample representativeness, validate measures of racial context, and assess whether AI-generated outputs accurately reflect the underlying population. When those benchmarks are themselves products of algorithmic processes, the external standard against which bias is measured becomes unreliable, and errors introduced into the research infrastructure may escape detection precisely because the tools for detecting them are subject to the same distortions.

Recent research raises concern that newly introduced privacy-preserving procedures in the Census inject noise into population counts in ways that obscure their accuracy. These distortions are especially pronounced for racial minorities, particularly Hispanic, Asian, and multiracial populations, and for smaller geographic units (Kenny et al., 2021; Kenny et al., 2024; Bozick et al., 2023). These inaccuracies can systematically distort the construction of core measures that scholars rely on, such as estimates of local racial context and segregation, benchmarks for minority political representation, and demographic baselines used to monitor racial equity and civil rights compliance (Mervis, 2024; Neidert et al., 2025; Asquith et al., 2022). When distorted benchmarks are then used to evaluate data sources and validate models that are themselves shaped by algorithmic processes, the result is a self-reinforcing cycle in which algorithmic errors are normalized rather than detected. Benchmarking data that are algorithmically constructed need to be evaluated as part of the measurement process itself, not assumed to provide external ground truth.

4.H AI, Agenda Setting, and Scholarly Knowledge Production

A final methodological concern operates at the level of the discipline itself. As political scientists increasingly use AI tools to summarize literatures, synthesize large text corpora, and identify dominant themes, these systems shape what appears central, representative, or theoretically salient (Delgado-Chaves et al., 2024; Bender et al., 2021; Weidinger et al., 2021; Grimmer et al., 2022). By design, large language models surface frequent patterns and smooth disagreement, privileging consensus and majority perspectives.

This dynamic poses particular risks for race and ethnic politics. Minority and marginalized perspectives are more likely to be compressed, sidelined, or categorized as outliers because they appear less frequently, use different linguistic registers, or challenge dominant narratives (Benjamin, 2019; Noble, 2018). Wagner, Lukyanenko, and Paré (2022) demonstrate that AI-assisted literature review tools structurally favor majority consensus positions, often suppressing contradictory perspectives. Over time, such compression may influence theory development, case selection, and the boundaries of legitimate inquiry, reshaping the scholarly canon under the appearance of neutrality. How AI-assisted research workflows shape whose ideas are treated as central and whose as marginal is therefore a necessary object of inquiry for a race-conscious methodology.

5. AI Measurement Risk and Disclosure Standards

Political science lacks shared standards for documenting how AI research infrastructure performs across socially relevant groups. The mechanisms documented in Section 4 – implicit racial inference, construct automation, differential translation quality, sampling exclusion, and benchmarking distortion – are structural features of AI systems that any researcher using these tools as measurement instruments may encounter. When that variation goes undocumented, measurement error enters the research record without a trace that peer review can evaluate or replication can detect. Without transparency about how AI outputs were produced, which populations were used to validate them, and where group-differentiated error entered the analysis, political scientists risk institutionalizing measurement artifacts as empirical fact.

5.A AI Measurement Statement (AIMS)

We propose an AI Measurement Statement (AIMS), a focused disclosure standard organized around four questions designed to surface group-differentiated measurement risks at the stages where AI systems most directly shape substantive inference. AIMS does not require exhaustive technical audits or specific performance thresholds. Its focus is transparency, making visible the measurement rules embedded in AI systems and the robustness of the empirical claims that depend on them. The framework is designed to be feasible for researchers without technical AI backgrounds and to apply across political science subfields. The depth of disclosure should be commensurate with the centrality of AI systems to the study's design and claims.

Core AIMS Components:

1. Instrument: What AI system was used, for what task, and where do its outputs enter the analysis?

Authors should specify the model or system employed, including provider, model name, and version or access date, along with the specific task performed, whether classification, prediction, translation, generation, or estimation. They should also indicate how the system's outputs function in the research design, as dependent variables, independent variables, preprocessing or filtering tools, or components of model estimation.

This disclosure establishes the properties of the measurement instrument and clarifies its inferential role. A classifier generating a dependent variable poses different validity risks than a system used only for preprocessing, and those risks operate differently across groups. A system that performs adequately on average may systematically perform worse on minority populations, and where that system generates the study’s core outcome measure, that error propagates directly into substantive conclusions. Because many proprietary and nondeterministic systems cannot meet traditional replication standards, AIMS follows Barrie, Palmer, and Spirling (2025) in emphasizing auditability over exact reproducibility, so readers can evaluate how results were generated even when they cannot reproduce them exactly.

2. Configuration: What choices shaped how the AI system operated, and what remains unobservable about how it processed key concepts?

Authors should document the decisions that directed the system, including prompts, instructions, or input specifications provided; preprocessing steps applied before the system received data; and the classification scheme or output categories the system produced. They should also document

how social categories such as race, ethnicity, gender, ideology, or language were constructed and implemented within the system. Where implementation details are unavailable because the system is proprietary, training data is undisclosed, or internal logic is opaque, that should be stated explicitly, along with what that opacity implies for evaluating construct validity.

This question addresses two related concerns. The first is the researcher's role in shaping how the tool operates. Most AI systems require configuration choices that directly shape outputs. Documenting those choices allows readers to evaluate whether the system was measuring what the researcher intended. The second concern is what cannot be known. For research involving socially relevant groups, opacity about training data composition and internal classification logic is a direct threat to construct validity. Systems trained on unrepresentative data may operationalize group concepts in ways that neither the researcher nor the reader can detect. Acknowledging both what was chosen and what remains hidden is essential to evaluating the measurement claims that follow.

3. Differential Performance: How might this AI system perform differently across socially relevant groups, and what are the inferential limits of the findings?

Authors should identify plausible sources of group-differentiated error, including differential accuracy, representational gaps in training data, domain mismatch, or systematic misclassification, and explain how these risks may distort substantive inferences. The disclosure should bound empirical claims accordingly, specifying the populations and conditions for which findings are most and least reliable.

A system that performs well on average may nonetheless bias group comparisons if error rates vary systematically across populations. Translation systems may privilege standardized language forms. Predictive models may exhibit higher error in marginalized communities. This question asks researchers to make those risks explicit and to specify for whom and under what conditions the findings hold.

4. Validation: How were AI-generated outputs evaluated for group-differentiated performance, and what materials support verification or replication?

Authors should describe validation procedures conducted on the system's outputs, with particular attention to subgroup-specific assessments where feasible, and they should report what those checks found. They should also indicate what materials have been archived, including prompts, coding scripts, sample outputs, validation data, and model version information, and note any practical limits on verification or replication.

Preserving documentation of system use and validation procedures enables meaningful peer evaluation even as underlying tools evolve. Where subgroup validation was infeasible, authors should clarify why and explain what that limitation implies for inference.

5.B Illustrative Examples of AIMS

The following examples show how AIMS applies across political science subfields and AI modalities, from text classification in comparative research to computer vision and machine translation in studies of minority political behavior. These represent best-practice disclosure.

AIMS is most valuable precisely in cases where researchers have not yet considered whether their AI tools perform differently across the populations they study.

Example 1: Text Classification in Comparative Research

1. Instrument: We used a multilingual LLM (mBERT, accessed June 2026) to classify, by populist rhetoric and ideological position, party manifestos from twelve Latin American countries. Classifications were used as independent variables in models of electoral volatility.

2. **Configuration**: We defined populism using a codebook developed from Spanish-language primary sources and applied consistent category labels across countries. Classification prompts were written in English and translated prior to application, introducing a layer of linguistic mediation not present for higher-resource European languages.

3. **Differential Performance**: Model performance was weaker for regional Spanish variants and for political vocabularies specific to particular national contexts. Classifications in countries with less-resourced training data representation may conflate ideological distinctions that human coders would treat as substantively distinct. Findings should be interpreted with caution for countries underrepresented in the underlying training corpus.

4. **Validation**: We assessed classification accuracy against a human-coded validation set stratified by country. We reported subgroup error rates by region. Prompts, model version, and validation data were archived. Because the model was accessed via API, future replication may encounter version differences.

Example 2: Computer Vision in Conflict and Mobilization Research

1. Instrument: We used Amazon Rekognition (accessed February 2025) to detect and identify faces in protest footage. These outputs were used to construct measures of protest participation and network ties.

2. Configuration: We relied on the system's built-in demographic classification categories based on facial features and skin tone to infer racial identity. These categories, thresholds, and confidence cutoffs were adopted from the default system configuration. The training data and algorithmic logic underlying these classifications are proprietary – meaning the specific features driving racial categorization cannot be independently evaluated.

3. Differential Performance: Prior audits indicate higher misidentification rates for darker-skinned individuals, increasing the risk of undercounting participation in racially diverse protests. These errors may bias estimates of mobilization and network centrality.

4. Validation: We compared automated identifications with hand-coded samples across racial groups and reported subgroup-specific error rates. Model version numbers, preprocessing scripts, and validation data were archived. System updates may affect future replication.

Example 3: AI-Administered Survey in Multilingual Political Behavior Research

1. **Instrument:** We used an AI conversational survey agent (built on GPT-4o, accessed March 2026) to administer open-ended interviews about political trust and civic

engagement with first- and second-generation immigrant respondents, in English, Spanish, and Vietnamese. Transcribed responses were coded to construct measures of institutional trust and political participation.

2. **Configuration:** The agent was designed to probe initial responses with follow-up questions adapted to the respondent's language. Persona design, question sequencing, and probing logic were developed by the research team and tested in pilot interviews. Respondents were informed that the interviewer was AI-generated. The system's internal decisions about when and how to probe, including tone, phrasing, and persistence, could not be fully observed or standardized across languages.
3. **Differential Performance:** The agent's natural language processing, probing logic, and response interpretation are likely to perform unevenly across the three languages. English-language capabilities in GPT-4o are substantially more developed than Vietnamese or Spanish capabilities. The system's ability to generate contextually appropriate follow-up questions, interpret idiomatic responses, and accurately code political meaning will reflect that gap. These technical disparities may compound with respondent-side effects: the disclosure that the interviewer is AI-generated may affect behavior differently across communities, particularly those with histories of state surveillance and institutional avoidance. Findings should be interpreted as most reliable for English-speaking, second-generation respondents. For first-generation respondents interviewed in Vietnamese and Spanish, both system-level language performance gaps and group-differentiated interviewer effects are plausible sources of measurement distortion.
4. **Validation:** Across language groups and against a matched sample of human-administered interviews, we compared response length, completion rates, and expressed political attitudes, assessing whether attitudinal patterns diverged systematically between interview modes within each group. These checks address surface-level response quality but do not capture variation in probing depth, interpretive accuracy, or conversational tone across languages. A fuller assessment would require bilingual expert review of transcripts against the agent's probing logic, which was beyond the scope of this study. Findings for Vietnamese- and Spanish-language interviews should be treated as provisional. Agent configuration files, interview transcripts, and coding protocols were archived.

5.C Relationship to Existing Standards

Political science has built strong norms around transparency and reproducibility. The American Political Science Association's Data Access and Research Transparency (DA-RT) framework requires researchers to disclose data, analytical procedures, and code to enable verification and cumulative inquiry. Leading journals have begun requiring disclosure of AI tool use (APSA, 2024). These standards represent genuine disciplinary achievements, but they were developed for conventional statistical models whose procedures are fully specifiable, parameters are observable, and outputs are deterministic given the same inputs. Contemporary AI systems have none of these properties. They are opaque, stochastic, dependent on proprietary training data, and capable of

producing systematically different outputs across demographic groups in ways that standard replication requirements do not surface. AIMS addresses that gap.

AIMS disclosures belong in methods sections, appendices, or replication materials, paralleling long-standing practices for documenting survey instruments, experimental protocols, and human coding procedures. AIMS extends these standards to a class of tools that increasingly function as measurement infrastructure in social science research. For researchers who already document measurement carefully, AIMS imposes minimal marginal burden. For those treating AI tools as self-explanatory, it codifies the documentation necessary to evaluate measurement quality and the claims that depend on it.

AIMS also reflects documentation practices emerging in computer science and industry. Model cards and system cards were introduced to standardize reporting of intended use, evaluation procedures, and subgroup variation (Mitchell et al., 2019; Mehraj et al., 2025). Commercial transparency reporting now includes documentation of evaluation suites, known limitations, and mitigation strategies (OpenAI, 2025). Governance frameworks increasingly treat such documentation as part of responsible AI integration (NIST, 2023, 2024). Independent audits continue to identify gaps, particularly in training data transparency and post-deployment monitoring (Wan et al., 2025), underscoring the value of user-side disclosure standards in downstream research.

The following table (Table 5.1) maps each AIMS question to analogous practices in both traditions.

Table 5.1. Mapping the AIMS four-question framework to existing methods standards

AIMS question	Purpose for inference	Analog in political science methods	Analog in industry practice
Instrument What AI system was used, for what task, and where do its outputs enter the analysis?	Clarifies what the measurement instrument is, how it functions in the analytic pipeline, and what validity risks follow from its inferential role.	Survey instrument description; coding scheme; software version reporting; variable construction documentation.	Model card "Intended Use;" system documentation; integration documentation.
Configuration What choices shaped how the AI system operated, and what remains	Defines the measurement rule and enables evaluation of construct validity.	Question wording; codebook definitions; coder training materials;	Prompt templates and labeling protocols; Model card "Training Data" and "Factors"

unobservable about how it processed key concepts?		documentation of missing or unavailable design information.	sections; system card design documentation.
Differential Performance How might this system perform differently across socially relevant groups, and what are the inferential limits of the findings?	Identifies group-differentiated error and bounds interpretability.	Measurement error discussion; subgroup robustness checks; documentation of scope conditions.	Disaggregated evaluation results; fairness audits; Model card "Quantitative Analyses" section.
Validation How were AI-generated outputs evaluated for group-differentiated performance, and what materials support verification or replication?	Supports auditability and informed peer review.	Intercoder reliability; validation samples; replication materials.	Evaluation suites; documentation artifacts; transparency reporting.

The most directly comparable framework is the GUIDE-LLM checklist developed by Feuerriegel et al. (2026) for large language model use in behavioral and social science research. GUIDE-LLM includes one optional item inviting researchers to "note any subgroup analyses." AIMS makes group-stratified documentation required and structures it across four questions tied to distinct stages of the research pipeline. A researcher who completes AIMS will satisfy that optional GUIDE-LLM item with far greater specificity and will address measurement risks that GUIDE-LLM's required items do not ask about.

The goal of AIMS is not to add a compliance requirement but to make this type of disclosure routine and, over time, to generate a cumulative record of how AI measurement tools perform across politically relevant groups. Three pathways can accelerate adoption. First, journals that already require AI disclosure should extend their submission guidelines to include group-differentiated performance reporting. Adding a brief AIMS-aligned requirement to existing checklists imposes minimal additional burden on authors while standardizing the information

reviewers need to evaluate AI-dependent findings. Second, APSA organized sections whose research regularly involves politically relevant group comparisons should establish AIMS as a subfield norm. Sections focused on race, ethnicity, and politics, gender and politics, migration and citizenship, and comparative political behavior are well positioned to signal expectations to authors before the journal submission stage, when documentation is easiest to produce. Third, funders supporting research at the intersection of AI and political inquiry should incorporate AIMS-aligned disclosure into grant reporting requirements. Many already expect responsible AI documentation as a condition of compliance. AIMS gives that expectation a concrete, researcher-facing structure. To reduce the documentation burden for first-time adopters, a community-maintained repository of annotated AIMS examples, organized by method and subfield, would serve both as a practical resource and as a cumulative record of how AI measurement tools perform across politically relevant populations. Over time, that record would allow the field to identify systematic limitations before they propagate through the literature.

AIMS is an initial framework. The four questions proposed here reflect the measurement risks most visible today, but the landscape of AI tools and their integration into political science research will continue to evolve. Scholars should refine these questions, expand the examples, and adapt the framework to research designs not anticipated here. The standard will be stronger for being built collaboratively across subfields.

6. Conclusion

This chapter has surveyed the landscape of AI's interaction with racial and ethnic politics across governance, participation, and scholarly inquiry, and it has identified a set of empirical, theoretical, and methodological gaps that demand sustained attention from the discipline. In public administration, algorithmic systems are altering the exercise of discretion, the scope of surveillance, and the distribution of administrative burden in ways that existing evidence suggests fall unevenly across racial and ethnic groups. In campaigns and political communication, AI is reshaping information environments, mobilization strategies, and organizational capacity through mechanisms whose downstream effects on minority political participation remain underspecified. In electoral administration, automated systems influence access, error distribution, and resource allocation at a scale that amplifies the consequences of differential performance. Across these domains, whether AI reproduces, reconfigures, or interrupts racialized political hierarchies depends on conditions that political science has not yet examined with sufficient empirical precision.

These dynamics are unlikely to operate in isolation. Fragmented information environments may undermine the coordination that collective action requires. Weakened organizational infrastructure increases vulnerability to administrative barriers. Rising participation costs erode the political capacity needed to contest the algorithmic systems producing those costs. If these feedback loops operate as the available evidence suggests, their cumulative effects on minority political power could be substantial. Yet material harm does not automatically generate political salience, and specifying the conditions under which AI-mediated harms become politically visible and actionable for affected communities remains an important question this chapter raises.

The methodological stakes are equally pressing. As AI tools are integrated into research workflows to classify text, operationalize constructs, administer surveys, build corpora, and synthesize literatures, they function as measurement instruments whose error properties shape inference and theory building across the discipline. The mechanisms documented in Section 4 are not peripheral concerns for methodological specialists. They bear on the reliability and cumulative character of political science research broadly, and they are most consequential in precisely the domains where valid measurement of group differences matters most. The AI Measurement Statement proposed in this chapter responds by formalizing four disclosure questions that allow peer reviewers and replicators to evaluate how AI outputs were produced and what they imply for inference. Its adoption would extend to AI tools the same documentation standards the discipline already applies to survey instruments, coding schemes, and experimental protocols.

While this chapter centers on the United States, the dynamics it identifies invite comparative inquiry. Similar AI systems encounter different structures of group hierarchy, state capacity, and institutional accountability across national contexts. The European Union's AI Act establishes risk-based regulatory categories and mandatory impact assessments that have no current equivalent in US federal law, and tracing how similar technical systems yield different political consequences across institutional settings would sharpen the questions this chapter raises. Comparative scholars are also well positioned to examine how AI interacts with ethnic and religious cleavage structures, caste hierarchies, and colonial legacies that produce analogous dynamics of algorithmic classification, differential performance, and political exclusion. The argument speaks with equal force to scholarship on gender and politics, where well-documented gender-differentiated performance in AI systems parallels the racial dynamics examined here and where intersectional analysis is essential for understanding compounded measurement error. More broadly, scholars of political communication, public opinion, and international relations who rely on AI tools for content analysis, survey processing, and automated event coding all face the measurement risks documented in this chapter. The AIMS framework is designed to be subfield-agnostic precisely because the underlying problem is structural.

For scholars of race and ethnic politics, AI represents a particularly rich and urgent site of inquiry, one where racial meaning, political membership, and power are being produced and contested under conditions of automation and scale. The subfield’s theoretical resources – including racial formation, Linked Fate, administrative burden, racialized surveillance, and epistemic power – are precisely what is needed to analyze systems whose political consequences are embedded in technical design choices that appear neutral. The discipline possesses the conceptual tools to meet this challenge. Whether it applies them before algorithmic outputs become the unexamined foundation of its empirical record is a question whose answer will shape the credibility of political science research for the foreseeable future.

7. Appendix: AI Measurement Statement (AIMS) Disclosure Template

The following template operationalizes the four-question AIMS framework as a practical disclosure instrument. Researchers should complete each section in proportion to the centrality of AI tools in their study design. A system that generates core outcome measures warrants fuller documentation than one used only for preprocessing.

1. What AI system was used, for what task, and where do its outputs enter the analysis?

Describe the AI tool or model, its version or access date, the specific task performed, and how its outputs function in the research design (e.g., as dependent variables, independent variables, data preprocessing or filtering tools, or as part of model estimation).

2. What choices shaped how the AI system operated, and what remains unobservable about how it processed key concepts?

Document the decisions that directed the system, including prompts, instructions, input specifications, preprocessing steps, and classification schemes. Describe how social categories such as race, ethnicity, gender, ideology, or language were constructed and implemented. Where implementation details are unavailable because the system is proprietary, training data is undisclosed, or internal logic is opaque, state that explicitly and note what it implies for evaluating construct validity.

3. How might this AI system perform differently across socially relevant groups, and what are the inferential limits of the findings?

Identify plausible sources of group-differentiated error, describe where and how accuracy or reliability differs across relevant populations, and explain how these differences may affect the interpretation of substantive findings. Specify the populations and conditions for which findings are most and least reliable.

4. How were AI-generated outputs evaluated for group-differentiated performance, and what materials support verification or replication?

Describe validation procedures conducted on the system's outputs, with particular attention to subgroup-specific assessments. Indicate what materials are archived or documented, including prompts, coding scripts, sample outputs, validation data, and model version information, and note any practical limits on verification or replication.

Reference

Agadjanian, Alexander. 2022. “How Many Americans Change Their Racial Identification over Time?” Socius: Sociological Research for a Dynamic World 8: 23780231221098547.

Alon-Barkat, Saar. 2025. “Algorithmic Discrimination in Public Service Provision.” *Journal of Public Administration Research and Theory* 35 (4): 469–486.

Altman, Micah, and Michael P. McDonald. 2010. “The Promise and Perils of Computers in Redistricting.” *Duke Journal of Constitutional Law & Public Policy* 5 (1): 69–111.

American Journal of Political Science. "Review Process." Accessed 2025. https://apsanet.org.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine Bias.” *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arias, Eric, Pablo Balán, Horacio Larreguy, John Marshall, and Pablo Querubín. 2019. “Information Provision, Voter Coordination, and Electoral Accountability: Evidence from Mexican Social Networks.” *American Political Science Review* 113 (2): 475–498.

Arnett, C. 2020. “Race, Surveillance, Resistance.” *Ohio State Law Journal* 81 (6): 1103–1142.

Asad, Asad L. "Latinos' Deportation Fears by Citizenship and Legal Status, 2007 to 2018." *Proceedings of the National Academy of Sciences* 117, no. 16 (2020): 8836–8844.

Asquith, Brian, Brad Hershbein, Tracy Kugler, Shane Reed, Steven Ruggles, Jonathan Schroeder, Steve Yesiltepe, and David Van Riper. 2022. “Assessing the Impact of Differential Privacy on Measures of Population and Racial Residential Segregation.” Harvard Data Science Review 4 (Special Issue 2).

Auxier, Brooke. "Activism on Social Media Varies by Race and Ethnicity, Age, Political Party." Washington, DC: Pew Research Center, July 13, 2020.

Baldus, David C., Charles Pulaski, and George Woodworth. 1983. “Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience.” Journal of Criminal Law and Criminology 74 (3): 661–753.

Barberá, Pablo, and Gonzalo Rivero. "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* 33, no. 6 (2015): 712–729. https://doi.org/10.1177/0894439314558836.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press.

Barocas, Solon, and Andrew D. Selbst. 2016. “Big Data’s Disparate Impact.” *California Law Review* 104 (3): 671–732.

Barreto, M., and G. M. Segura. 2014. *Latino America: How America's Most Dynamic Population Is Poised to Transform the Politics of the Nation*. PublicAffairs.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace. 2022. "Consumer-Lending Discrimination in the FinTech Era." *Journal of Financial Economics* 143(1): 30–56

Becker, Amariah, Moon Duchin, Dara Gold, and Sam Hirsch. 2021. "Computational Redistricting and the Voting Rights Act." *Election Law Journal* 20 (4): 407–441.

Behrens, Angela, Christopher Uggen, and Jeff Manza. 2003. "Ballot Manipulation and the 'Menace of Negro Domination': Racial Threat and Felon Disenfranchisement in the United States, 1850–2002." *American Journal of Sociology* 109 (3): 559–605.

Benjamin, A. 2017. *Racial Coalition Building in Local Elections: Elite Cues and Cross-Ethnic Voting*. Cambridge University Press.

Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Benjamin, Ruha. 2023. "Race after Technology." In *Social Theory Re-Wired*, 405–415. Routledge.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bennett Moses, L., and J. Chan. 2018. "Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability." *Policing and Society* 28 (7): 806–822.

Benthall, Sebastian, and Bruce D. Haynes. 2019. "Racial Categories in Machine Learning." In Proceedings of the Conference on Fairness, Accountability, and Transparency, 289–298.

Blank, Grant. 2017. "The Digital Divide among Twitter Users and Its Implications for Social Research." *Social Science Computer Review* 35 (6): 679–697.

Blodgett, Su Lin, and Brendan O'Connor. 2017. "Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English." *arXiv* preprint arXiv:1707.00061.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP." *arXiv* preprint arXiv:2005.14050.

Blumenstein, R. 2021. "The Perfect Match: Solving the Due Process Problem of Signature Matching with Federal Agency Regulation." *Vanderbilt Journal of Entertainment & Technology Law* 24 (1): 121–156.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS 2016)*, 4356–4364. Barcelona, Spain: Curran Associates, Inc.

Bozick, Robert, Lane F. Burgette, Ethan Sharygin, Regina A. Shih, Beverly Weidmer, Michael Tzen, Aaron Kofner, Jennie E. Brand, and Hiram Beltrán-Sánchez. 2023. “Evaluating the Accuracy of 2020 Census Block-Level Estimates in California.” Demography 60 (6): 1903–1921.

Brand, Jennie E., Xiang Zhou, and Yu Xie. 2023. “Recent Developments in Causal Inference and Machine Learning.” Annual Review of Sociology 49 (1): 81–110.

Brayne, Sarah. 2017. “Big Data Surveillance: The Case of Policing.” *American Sociological Review* 82 (5): 977–1008.

Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of the Conference on Fairness, Accountability and Transparency*, 77–91. PMLR.

Burrell, Jenna. 2016. “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.” *Big Data & Society* 3 (1): 2053951715622512.

Cable, J., A. Fábrega, S. Park, and M. Specter. 2023. “A Systematization of Voter Registration Security.” *Journal of Cybersecurity* 9 (1): Article tyad008.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.” *Science* 356 (6334): 183–186.

Canon, David T. 2022. “Race and Redistricting.” *Annual Review of Political Science* 25: 509–528.

Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback.” *Transactions on Machine Learning Research* (accepted December 30, 2023). Preprint, arXiv:2307.1521.

Chen, Jowei, and Jonathan Rodden. 2013. "Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures." *Quarterly Journal of Political Science* 8(3): 239–269.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. “Double/Debiased Machine Learning for Treatment and Structural Parameters.” Econometrics Journal 21 (1): C1–C68.

Cho, Wendy Tam, and Bruce E. Cain. 2020. “Human-Centered Redistricting Automation in the Age of AI.” *Science* 369 (6508): 1179–1181.

Chong, Dennis. 1991. *Collective Action and the Civil Rights Movement*. University of Chicago Press.

Chong, Dennis, and Reuel Rogers. 2005. “Racial Solidarity and Political Participation.” *Political Behavior* 27 (4): 347–374.

Chouldechova, Alexandra. 2017. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data* 5 (2): 153–163.

Cohen, Cathy J. 1999. *The Boundaries of Blackness: AIDS and the Breakdown of Black Politics*. University of Chicago Press.

Coltin, Jeff. 2023. “Greetings from Mayor Adams, generated by AI, in different languages.” *Politico*, October 16, https://www.politico.com/news/2023/10/16/nyc-adams-ai-languages-00121744

Cox, Gary W., and Mathew D. McCubbins. 1986. “Electoral Politics as a Redistributive Game.” *The Journal of Politics* 48 (2): 370–389.

Cranmer, Skyler J. 2019. “Introduction to the Virtual Issue: Machine Learning in Political Science.” Political Analysis 27 (1): 1–9.

DaViera, Andrea L., Marbella Uriostegui, Aaron Gottlieb, and Ogechi Onyeka. 2024. "Risk, race, and predictive policing: A critical race theory analysis of the strategic subject list." *American journal of community psychology* 73: 91-103.

Davidson, Thomas, Debasmita Bhattacharya, and Ingmar Weber. 2019. “Racial Bias in Hate Speech and Abusive Language Detection Datasets.” *arXiv* preprint arXiv:1905.12516.

Dawson, Michael C. 1994. *Behind the Mule: Race and Class in African-American Politics*. Princeton University Press.

Delgado-Chaves, Fernando M., Matthew J. Jennings, Antonio Atalaia, Justus Wolff, Rita Horvath, Zeinab M. Mamdouh, Jan Baumbach, and Linda Baumbach. 2025. “Transforming Literature Screening: The Emerging Role of Large Language Models in Systematic Reviews.” *Proceedings of the National Academy of Sciences* 122 (2): e2411962122.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus.” *arXiv* (April). arXiv:2104.08758.

Dong, E., A. Schein, Y. Wang, and N. Garg. 2025. “Addressing Discretization-Induced Bias in Demographic Prediction.” *PNAS Nexus* 4 (2): pgaf027.

Dorie, Vincent, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 2019. “Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition.” Statistical Science 34 (1): 43–68.

Dressel, Julia, and Hany Farid. 2018. “The Accuracy, Fairness, and Limits of Predicting Recidivism.” Science Advances 4 (1): eaao5580.

Eagly, Ingrid V. 2017. “Criminal Justice in an Era of Mass Deportation: Reforms from California.” *New Criminal Law Review* 20 (1): 12–38.

Egami, Naoki; Musashi Hinck, Brandon Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36: 68589–68601, 2023.

Egami, Naoki; Musashi Hinck, Brandon M Stewart, and Hanying Wei. Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses. 2024.

Elazar, Yanai, and Yoav Goldberg. 2018. “Adversarial Removal of Demographic Attributes from Text Data.” *arXiv* preprint arXiv:1808.06640.

Endres, K., and K. J. Kelly. 2018. “Does Microtargeting Matter? Campaign Contact Strategies and Young Voters.” *Journal of Elections, Public Opinion and Parties* 28 (1): 1–18.

Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. “Runaway Feedback Loops in Predictive Policing.” *Proceedings of Machine Learning Research* 81: 160–176.

Epp, Charles R., Steven Maynard-Moody, and Donald Haider-Markel. 2014. *Pulled Over: How Police Stops Define Race and Citizenship*. University of Chicago Press.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin’s Press.

Farzan, Antonia Noori. 2018. “Memphis Police Used a Fake Facebook Account to Monitor Black Lives Matter, Trial Reveals.” *The Washington Post*, August 23.

Feuerriegel, Stefan, Christopher Barrie, M. J. Crockett, Laura K. Globig, Killian L. McLoughlin, Dan-Mircea Mirea, Manoel Horta Ribeiro, Steve Rathje, Arthur Spirling, and Diyi Yang. 2026. "GUIDE-LLM: A Consensus-Based Reporting Checklist for Large Language Models in Behavioral and Social Science." URL https://sfeuerriegel.github.io/llm-checklist/. Accessed 21 February 2026.

Fiorina, Morris P. 1981. *Retrospective Voting in American National Elections*. Yale University Press.

Fleisig, Eve, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. “Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination.” In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13541–13564. Miami, FL: Association for Computational Linguistics.

Flores, A., and A. Coppock. 2018. “Do Bilinguals Respond More Favorably to Candidate Advertisements in English or in Spanish?” *Political Communication* 35 (4): 612–633.

Fraga, Bernard L. 2018. *The Turnout Gap: Race, Ethnicity, and Political Inequality in a Diversifying America*. Cambridge University Press.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. “The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making.” *Communications of the ACM* 64 (4): 136–143. https://doi.org/10.1145/3433949.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably Unequal? The Effects of Machine Learning on Credit Markets." *Journal of Finance* 77(1): 5–47).

Garcia, Jennifer, and Christopher Stout. 2022. “The Empowering Effects of Racial Messaging: The Link between Racial Outreach, Descriptive Representation and Black Political Mobilization.” *Political Communication* 39 (5): 589–606.

Garcia-Bedolla, Lisa, and Melissa R. Michelson. 2012. *Mobilizing Inclusion: Transforming the Electorate through Get-Out-the-Vote Campaigns*. Yale University Press.

Gay, Claudine, Jennifer Hochschild, and Ariel White. 2016. "Americans' Belief in Linked Fate: Does the Measure Capture the Concept?" *Journal of Race, Ethnicity, and Politics* 1(1): 117-144

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.

Gilens, Martin. 1999. *Why Americans Hate Welfare: Race, Media, and the Politics of Antipoverty Policy*. University of Chicago Press.

Glaser, James M. 1996. *Race, Campaign Politics, and the Realignment in the South*. New Haven: Yale University Press.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Gray, T. R., and J. A. Jenkins. 2025. “Estimating Disenfranchisement in US Elections, 1870–1970.” *Perspectives on Politics* 23 (1): 55–75.

Green, Donald P., and Alan S. Gerber. 2019. *Get Out the Vote: How to Increase Voter Turnout*. Brookings Institution Press.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. “Machine Learning for Social Science: An Agnostic Approach.” *Annual Review of Political Science* 24 (1): 395–419.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grofman, Bernard, Lisa Handley, and Richard G. Niemi. 1992. *Minority Representation and the Quest for Voting Equality*. Cambridge University Press.

Guo, Yanzhu, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. "Do Large Language Models Have an English 'Accent'? Evaluating and Improving the Naturalness of Multilingual LLMs." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics.

Guerreiro, Nuno M., Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. "Hallucinations in Large Multilingual Translation Models." *Transactions of the Association for Computational Linguistics* 11: 1500–1517. https://doi.org/10.1162/tacl_a_00615.

Haimson, Oliver L., Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. "Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): Article 466.

Han, Hahrie. 2014. *How Organizations Develop Activists: Civic Associations and Leadership in the 21st Century*. Oxford University Press.

Hero, Rodney E. 1992. *Latinos and the US Political System: Two-Tiered Pluralism*. Temple University Press.

Hersh, Eitan D. 2015. *Hacking the Electorate: How Campaigns Perceive Voters*. Cambridge University Press.

Ho, Annabell, Jeff Hancock, and Adam S. Miner. 2018. "Psychological, Relational, and Emotional Effects of Self-Disclosure after Conversations with a Chatbot." *Journal of Communication* 68 (4): 712–733.

Hovy, Dirk, and Shrimai Prabhumoye. 2021. "Five sources of bias in natural language processing." *Language and linguistics compass* 15: e12432.

Hovy, Dirk, and Shannon L. Spruit. 2016. "The Social Impact of Natural Language Processing." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Huber, Gregory A., et al. 2021. "The Racial Burden of Voter List Maintenance Errors: Evidence from Wisconsin's Supplemental Movers Poll Books." *Science Advances* 7: eabe4498.

Humber, Nadiyah J. 2023. "A Home for Digital Equity: Algorithmic Redlining and Property Technology." *California Law Review* 111: 1103–1164

Huq, Aziz Z. 2019. "Racial Equity in Algorithmic Criminal Justice." *Duke Law Journal 68*: 1043-1134.

Hutchings, Vincent L., and Nicholas A. Valentino. 2004. “The Centrality of Race in American Politics.” *Annual Review of Political Science* 7: 383–408.

Imai, Kosuke, and Kabir Khanna. 2016. “Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records.” Political Analysis 24 (2): 263–272.

Imana, Basileal, Aleksandra Korolova, and John Heidemann. 2021. “Auditing for Discrimination in Algorithms Delivering Job Ads.” In *Proceedings of the Web Conference 2021 (WWW ’21)*, 3767–3778. New York: Association for Computing Machinery. https://doi.org/10.1145/3442381.3450077.

Jacobs, Abigail Z., and Hanna Wallach. 2021. “Measurement and Fairness.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

Jacobson, Gary C. 1989. “Strategic Politicians and the Dynamics of U.S. House Elections, 1946–86.” *American Political Science Review* 83 (3): 773–793.

Johnson, D., D. B. Wilson, E. R. Maguire, and B. V. Lowrey-Kinberg. 2017. “Race and Perceptions of Police: Experimental Results on the Impact of Procedural (In)Justice.” *Justice Quarterly* 34 (7): 1184–1212.

Jun, S., R. M. Chow, A. M. van der Veen, and E. Bleich. 2022. “Chronic Frames of Social Inequality: How Mainstream Media Frame Race, Gender, and Wealth Inequality.” *Proceedings of the National Academy of Sciences* 119 (21): e2110712119.

Junn, Jane, and Natalie Masuoka. 2008. “Asian American Identity: Shared Racial Status and Political Context.” *Perspectives on Politics* 6 (4): 729–740.

Jurka, Timothy P., Loren Collingwood, Amber E. Boydstun, and Emiliano Grossman. 2013. “RTextTools: A Supervised Learning Package for Text Classification.” *The R Journal* 5 (1): 6–12.

Kalla, Joshua L., and David E. Broockman. 2018. “The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments.” *American Political Science Review* 112 (1): 148–166.

Kärkkäinen, Kimmo, and Jungseock Joo. 2021. “FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation.” In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1548–1558.

Kasy, Maximilian. 2024. “Algorithmic Bias and Distributional Justice.” *Oxford Review of Economic Policy* 40 (3): 530–552.

Kaufmann, Karen M. 2003. “Cracks in the Rainbow: Group Commonality as a Basis for Latino and African-American Political Coalitions.” *Political Research Quarterly* 56 (2): 199–210.

Kenny, Christopher T., Shiro Kuriwaki, Cory McCartan, Evan T. R. Rosenman, Tyler Simko, and Kosuke Imai. 2021. "The Use of Differential Privacy for Census Data and Its Impact on Redistricting: The Case of the 2020 U.S. Census." Science Advances 7 (41).

Kenny, Christopher T., Cory McCartan, Shiro Kuriwaki, Tyler Simko, and Kosuke Imai. 2024. "Evaluating Bias and Noise Induced by the U.S. Census Bureau's Privacy Protection Methods." Science Advances 10 (18).

Keyssar, Alexander. 2009. *The Right to Vote: The Contested History of Democracy in the United States*. Basic Books.

Kim, Pauline T. 2022. "Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action." *California Law Review 110*: 1539-1596.

Kinder, Donald R., and Lynn M. Sanders. 1996. *Divided by Color: Racial Politics and Democratic Ideals*. University of Chicago Press.

Kiritchenko, Svetlana, and Saif Mohammad. 2018. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems." In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent Trade-Offs in the Fair Determination of Risk Scores." In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 43:1–43:23. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–293.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior." *Proceedings of the National Academy of Sciences* 110 (15): 5802–5805.

Kousser, J. Morgan. 1971. *The Shaping of Southern Politics: Suffrage Restriction and the Establishment of the One-Party South*. Yale University Press.

Kroll, Joshua A., Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (3): 633–705.

Kruis, N. E., R. H. Donohue, N. Glunt, N. J. Rowland, and J. Choi. 2023. "Examining the Effects of Perceptions of Police Effectiveness, Procedural Justice, and Legitimacy on Racial Differences in Anticipated Cooperation with Law Enforcement in Pennsylvania." *Criminal Justice Policy Review* 34 (8): 841–869.

Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. “Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning.” Proceedings of the National Academy of Sciences 116 (10): 4156–4165.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2010. “Improving Propensity Score Weighting Using Machine Learning.” Statistics in Medicine 29 (3): 337–346.

Lee, C., K. Gligorić, P. R. Kalluri, M. Harrington, E. Durmus, K. L. Sanchez, N. San, D. Tse, X. Zhao, M. G. Hamedani, and H. R. Markus. 2024. “People Who Share Encounters with Racism Are Silenced Online by Humans and Machines, but a Guideline-Reframing Intervention Holds Promise.” *Proceedings of the National Academy of Sciences* 121 (38): e2322764121.

Lee, Erika. 2019. *America for Americans: A History of Xenophobia in the United States*. Basic Books.

Lerman, Amy E., and Vesla M. Weaver. 2014. *Arresting Citizenship: The Democratic Consequences of American Crime Control*. University of Chicago Press.

Lerman, Amy E., and Vesla M. Weaver. "Political Consequences of the Carceral State." *American Political Science Review* 104, no. 4 (2010): 817–833.

Levitt, Justin. 2017. “Race, Redistricting, and the Manufactured Conundrum.” *Loyola of Los Angeles Law Review* 50: 555.

Lieberman, Robert C. 2001. *Shifting the Color Line: Race and the American Welfare State*. Harvard University Press.

Lum, Kristian, and William Isaac. 2016. “To Predict and Serve?” *Significance* 13 (5): 14–19.

Mauk, Marlene, and Max Grömping. 2024. “Online Disinformation Predicts Inaccurate Beliefs about Election Fairness among Both Winners and Losers.” *Comparative Political Studies* 57 (6): 965–998.

Mayson, Sandra G. 2019. “Bias in, Bias Out.” *Yale Law Journal* 128: 2218–2300.

McCartan, Cory, and Kosuke Imai. 2023. "Sequential Monte Carlo for Sampling Balanced and Compact Redistricting Plans." *Annals of Applied Statistics* 17(4): 3300–3323.

McCartan, C., C. T. Kenny, T. Simko, G. Garcia III, K. Wang, M. Wu, S. Kuriwaki, and K. Imai. 2022. “Simulated Redistricting Plans for the Analysis and Evaluation of Redistricting in the United States.” *Scientific Data* 9 (1): Article 689.

McNamara, R. G., and P. Tikka. 2023. “Well-Founded Fear of Algorithms or Algorithms of Well-Founded Fear? Hybrid Intelligence in Automated Asylum Seeker Interviews.” *Journal of Refugee Studies* 36 (2): 238–270.

Mehraj, Ali, An Cao, Kari Systä, Tommi Mikkonen, Pyry Kotilainen, David Hästbacka, and Niko Mäkitalo. 2025. “AI Model Cards: State of the Art and Path to Automated Use.” In *WEBIST*. SCITEPRESS Science and Technology Publications.

Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton University Press.

Meng, Amanda and Carl DiSalvo. 2018. “Grassroots Resource Mobilization through Counter-Data Action.” *Big Data & Society,* July-December: 1-12.

Mervis, Jeffrey. 2024. “The U.S. Has a New Way to Mask Census Data in the Name of Privacy. How Does It Affect Accuracy?” ScienceInsider. https://www.science.org/content/article/u-s-has-new-way-mask-census-data-name-privacy-how-does-it-affect-accuracy.

Mettler, Suzanne. 2011. *The Submerged State: How Invisible Government Policies Undermine American Democracy*. University of Chicago Press.

Mettler, Suzanne, and Joe Soss. 2004. “The Consequences of Public Policy for Democratic Citizenship: Bridging Policy Studies and Mass Politics.” *Perspectives on Politics* 2 (1): 55–73.

Michener, Jamila. 2018. *Fragmented Democracy: Medicaid, Federalism, and Unequal Politics*. Cambridge University Press.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. “Model Cards for Model Reporting.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Molnar, Petra. 2019. “Technology on the Margins: AI and Global Migration Management from a Human Rights Perspective.” *Cambridge International Law Journal* 8 (2): 305–330.

Molnar, Petra. 2024. *The Walls Have Eyes: Surviving Migration in the Age of Artificial Intelligence*. The New Press.

National Institute of Standards and Technology. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: U.S. Department of Commerce. January. https://doi.org/10.6028/NIST.AI.100-1.

National Institute of Standards and Technology. 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. Gaithersburg, MD: U.S. Department of Commerce. July. https://doi.org/10.6028/NIST.AI.600-1.

Neidert, Lisa, Reynolds Farley, and Jeffrey Morenoff. 2025. “How Census Undercount Became a Civil Rights Issue and Why It Is Increasingly Important.” RSF: The Russell Sage Foundation Journal of the Social Sciences 11 (1): 26–43.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366 (6464): 447–453.

Oh, D., and J. Downey. 2025. “Does Algorithmic Content Moderation Promote Democratic Discourse? Radical Democratic Critique of Toxic Language AI.” *Information, Communication & Society* 28 (7): 1157–1176.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." *Frontiers in Big Data* 2 (2019): 13. https://doi.org/10.3389/fdata.2019.00013.

Omi, Michael, and Howard Winant. 1986. *Racial Formation in the United States*. Routledge.

OpenAI. 2025. *OpenAI GPT-4.5 System Card*. February 27. https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf.

Overton, S. 2024. “Overcoming Racial Harms to Democracy from Artificial Intelligence.” *Iowa Law Review* 110: 805.

Overton, Spencer A. 2026. "Ethnonationalism by Algorithm." Unpublished manuscript.

Palmer, Alexis, Noah A. Smith, and Arthur Spirling. 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.

Panditharatne, M. 2024. “Preparing to Fight AI-Backed Voter Suppression.” Brennan Center for Justice, April 16.

Papneja, Hashai, and Nikhil Yadav. 2025. “Self-Disclosure to Conversational AI: A Literature Review, Emergent Framework, and Directions for Future Research.” *Personal and Ubiquitous Computing* 29 (2): 119–151.

Penner, Andrew M. and Aliya Saperstein. 2008. “How Social Status Shapes Race.” *PNAS 105*: 19628–19630.

Penney, Jonathon W. 2016. “Chilling Effects: Online Surveillance and Wikipedia Use.” *Berkeley Technology Law Journal* 31: 117.

Penney, Jonathon W. "Understanding Chilling Effects." *Minnesota Law Review* 106, no. 3 (2022): 1451–1530.

Pedraza, Francisco I., Victoria C. Nichols, and Adriana M. LeBrón. 2017. “Cautious Citizenship: The Deterring Effect of Immigration Issue Salience on Health Care Use and Bureaucratic Interactions among Latino US Citizens.” *Journal of Health Politics, Policy and Law* 42 (5): 925–960.

Piccardi, Tiziano, Martin Saveski, Chenyan Jia, Jeffrey Hancock, Jeanne L. Tsai, and Michael S. Bernstein. 2025. "Reranking partisan animosity in algorithmic social media feeds alters affective polarization." *Science* 390, no. 6776: eadu5584.

Pierson, Paul. 1993. "When Effect Becomes Cause: Policy Feedback and Political Change." *World Politics* 45 (4): 595–628.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. "On Fairness and Calibration." In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, 5684–5693. Long Beach, CA: Curran Associates, Inc.

Provine, Doris Marie. 2013. "Institutional Racism in Enforcing Immigration Law." *Norteamérica* 8: 31–53.

Qin, Libo, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. "A Survey of Multilingual Large Language Models." *Patterns* 6 (1) (January): 101118. https://doi.org/10.1016/j.patter.2024.101118.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time." Proceedings of the National Academy of Sciences 114 (41): 10870–10875.

Raji, Inioluwa Deborah, and Joy Buolamwini. 2019. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Ramakrishnan, S. K. 2006. *Democracy in Immigrant America: Changing Demographics and Political Participation*. Stanford University Press.

Ray, Victor, Pamela Herd, and Donald Moynihan. 2022. "Racialized Burdens: Applying Racialized Organization Theory to the Administrative State." *Journal of Public Administration Research and Theory* 33 (1): 139–152.

Rinaldi, Alberto, and Sue Anne Teo. 2025. "The Use of Artificial Intelligence Technologies in Border and Migration Control and the Subtle Erosion of Human Rights." *International and Comparative Law Quarterly* 74: 61–89.

Rosa, Jonathan and Nelson Flores. 2017. "Unsettling race and language: Toward a raciolinguistic perspective." *Language in Society* 46(5):621-647.

Ross, B. L. 2021. "Voter Data, Democratic Inequality, and the Risk of Political Violence." *Cornell Law Review* 107: 1011.

Ruths, Derek, and Jürgen Pfeffer. "Social Media for Large Studies of Behavior." *Science* 346, no. 6213 (November 28, 2014): 1063–1064. https://doi.org/10.1126/science.346.6213.1063.

Sanchez, Gabriel R. 2006. “The Role of Group Consciousness in Latino Public Opinion.” *Political Research Quarterly* 59 (3): 435–446.

Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2019. “The Risk of Racial Bias in Hate Speech Detection.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics. https://aclanthology.org/P19-1163/.

Savaget, P., T. Chiarini, and S. Evans. 2019. “Empowering Political Participation through Artificial Intelligence.” *Science and Public Policy* 46 (3): 369–380.

Schuetzler, Ryan M., Justin Scott Giboney, G. Mark Grimes, and Jay F. Nunamaker Jr. 2018. “The Influence of Conversational Agent Embodiment and Conversational Relevance on Socially Desirable Responding.” *Decision Support Systems* 114: 94–102.

Selbst, Andrew D., danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. “Fairness and Abstraction in Sociotechnical Systems.” In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* ’19)*, 59–68. New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598.

Sen, Maya, and Omar Wasow. 2016. “Race as a Bundle of Sticks: Designs That Estimate Effects of Seemingly Immutable Characteristics.” Annual Review of Political Science 19 (1): 499–522

Shah, Paru, and Robert S. Smith. 2021. “Legacies of Segregation and Disenfranchisement: The Road from *Plessy* to *Frank* and Voter ID Laws in the United States.” *The Russell Sage Foundation Journal of the Social Sciences* 7: 134–146.

Sinclair, Betsy, Margaret McConnell, and Melissa R. Michelson. 2013. “Local Canvassing: The Efficacy of Grassroots Voter Mobilization.” *Political Communication* 30 (1): 42–57.

Skeem, Jennifer L., and Christopher T. Lowenkamp. 2016. “Risk, Race, and Recidivism: Predictive Bias and Disparate Impact.” *Criminology* 54 (4): 680–712.

Soss, Joe. 1999. “Lessons of Welfare: Policy Design, Political Learning, and Political Action.” *American Political Science Review* 93 (2): 363–380.

Stewart, Charles. 2013. “Waiting to Vote in 2012.” *Journal of Law and Politics* 28: 439–463.

Sweeney, Latanya. 2013. “Discrimination in Online Ad Delivery.” *Communications of the ACM* 56 (5): 44–54.

Tate, Katherine. 1994. *From Protest to Politics: The New Black Voters in American Elections*. Harvard University Press.

Tesler, Michael. 2016. *Post-Racial or Most-Racial? Race and Politics in the Obama Era*. University of Chicago Press.

Timmons, Stephen, et al. 2022. “A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence.” *Journal of Community Health* 47 (6): 1021–1028.

Tyler, Tom R. 1990. *Why People Obey the Law*. New Haven, CT: Yale University Press.

Tyler, T. R., J. D. Casper, and B. Fisher. 1989. “Maintaining Allegiance toward Political Authorities: The Role of Prior Attitudes and the Use of Fair Procedures.” *American Journal of Political Science* 33 (3): 629–652.

Uribe, L., K. Aldridge, T. Kousser, K. Nichols-Smith, and T. Rush. 2025. “The Racial Gap in Trust in Elections (and How to Close It).” *Political Research Quarterly* 78 (4): 1408–1428.

VanderWeele, Tyler J., and Miguel A. Hernán. 2012. “Causal Effects and Natural Laws: Toward a Conceptualization of Causal Counterfactuals for Nonmanipulable Exposures, with Application to the Effects of Race and Sex.” In Causality: Statistical Perspectives and Applications, 101–113. Hoboken, NJ: Wiley.

Vargas, Edward D., Gabriel R. Sanchez, and Melina Juárez. "Fear by Association: Perceptions of Anti-Immigrant Policy and Health Outcomes." *Journal of Health Politics, Policy and Law* 42, no. 3 (2017): 459–483.

Varsanyi, Monica W. 2008. “Rescaling the ‘Alien,’ Rescaling Personhood: Neoliberalism, Immigration, and the State.” *Annals of the Association of American Geographers* 98 (4): 877–896.

Varsanyi, Monica W., Paul G. Lewis, Doris M. Provine, and Scott H. Decker. 2012. “A Multilayered Jurisdictional Patchwork: Immigration Federalism in the United States.” *Law & Policy* 34 (2): 138–158.

Wagner, Gerit, Roman Lukyanenko, and Guy Paré. "Artificial Intelligence and the Conduct of Literature Reviews." *Journal of Information Technology* 37, no. 2 (2022): 209–226. https://doi.org/10.1177/02683962211048201

Wan, Alexander, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, and Rishi Bommasani. 2025. *The 2025 Foundation Model Transparency Index*. Stanford Center for Research on Foundation Models (CRFM), December. https://crfm.stanford.edu/fmti/December-2025/paper.pdf.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. 2021. “Ethical and Social Risks of Harm from Language Models.” *arXiv* preprint arXiv:2112.04359.

White, Ariel. 2016. “When Threat Mobilizes: Immigration Enforcement and Latino Voter Turnout.” *Journal of Politics* 78 (4): 1137–1152.

Wilson, Kyra, and Aylin Caliskan. 2024. “Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval.” In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)*. Seattle, WA: Association for the Advancement of Artificial Intelligence.

Wolfe, Robert, Mahzarin R. Banaji, and Aylin Caliskan. 2022. “Evidence for Hypodescent in Visual Semantic AI.” In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1293–1304.

Wong, J. 2006. *Democracy’s Promise: Immigrants and American Civic Institutions*. University of Michigan Press.

Wuttke, Alexander, Matthias Aßenmacher, Christopher Klamm, Max Lang, and Fraue Kreuter. 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 (LaTeCH-CLfL 2025)*, 179–204.

Xiao, Jiancong, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. 2024. “On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization.” *arXiv* (May). arXiv:2405.16455.

Xiao, Ziang, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. “Tell Me about Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions.” *ACM Transactions on Computer-Human Interaction (TOCHI)* 27 (3): 1–37.

Yucer, Seyma, Furkan Tektas, Noura Al Moubayed, and Toby Breckon. 2024. “Racial Bias within Face Recognition: A Survey.” *ACM Computing Surveys* 57 (4): 1–39.

Zepeda-Millán, Chris. 2017. *Latino Mass Mobilization: Immigration, Racialization, and Activism*. Cambridge University Press.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. “Mitigating Unwanted Biases with Adversarial Learning.” In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods.” *arXiv* (April). arXiv:1804.06876.

Zhou, Xiang, and Guanghui Pan. 2023. “Higher Education and the Black-White Earnings Gap.” American Sociological Review 88 (1): 154–188.

AI statement

The authors used AI-assisted tools at limited stages of the research and writing process to support drafting, organization, and synthesis of material. These tools were used as aids for brainstorming, outlining, and language refinement. They were not used to collect, classify, code, or analyze data, and they did not generate any of the empirical claims, interpretations, or citations in this chapter. All arguments and conclusions were developed, verified, and revised by the authors through independent review of the relevant literature.